Kavli HUMAN Project

Preliminary Study Design

Kavli HUMAN Project

Preliminary Study Design Report

NEW YORK UNIVERSITY
Institute for the Interdisciplinary Study of Decision Making
October 15, 2015

Kavli HUMAN Project
Brooklyn, NY
www.kavlihumanproject.org

Institute for the Interdisciplinary Study of Decision Making
New York University
300 Cadman Plaza West, 7th Floor
Brooklyn, NY 11201

NOTICE: The Kavli HUMAN Project operates out of the New York University Institute for the Interdisciplinary Study of Decision Making (NYU IISDM) in partnership with The Kavli Foundation. This Preliminary Study Design was developed by staff of the Kavli HUMAN Project housed in NYU IISDM with support from The Kavli Foundation.

Edited by:
Kathleen R. Flaherty, Kavli HUMAN Project
Lewis B. Groswald, Kavli HUMAN Project

Cover Design by:
Echo Wang, NYU

Printed in the United States of America.

"The genetic analysis of neuropsychiatric disorders and neurodegenerative disorders is now advancing rapidly, based on new genomic technologies and computation. While the central goal of this research is to understand the biology of disease, an early discovery is that the genes predisposing to such neurodevelopmental disorders as autism spectrum disorders (ASDs), attention deficit hyperactivity disorder, and even schizophrenia, are normally distributed in populations. Illness occurs as a result of possessing a large enough number of risk alleles plus environmental risk factors. It also appears that in smaller numbers, the same alleles that give risk to ASDs may contribute to specific patterns of social cognition and behavior in unaffected individuals. What we are missing is deep phenotyping that would facilitate understanding of what specific combinations of genes signify for human cognition, emotion, and behavior."

The proposed Kavli HUMAN Project cohort has significant advantages for helping turn genetic discoveries into neurobiological information. Most important, in my view, is that they will learn how to deeply phenotype individuals and their results will form the basis for briefer, cheaper batteries to be employed in other studies. Second, they will have the possibility of stratifying individuals by genotype in order to study environment interactions effectively--current candidate methodologies in small samples have yielded no durable results.

In short, as a neurobiologist I am enthusiastic about this project as it will create a dataset that will make possible a great deal of mechanistic investigation."

- **Steven Hyman, MD**
Director, Stanley Center for Psychiatric Research at the Broad Institute
Former Director, United States National Institute for Mental Health (NIMH)
Former Provost, Harvard University

BOARD OF DIRECTORS

PAUL W. GLIMCHER, *Director*, Kavli HUMAN Project; *Julius Silver Professor* of Neural Science, Economics and Psychology; *Director*, Institute for the Interdisciplinary Study of Decision Making, New York University

ANDREW CAPLIN, *Chair*, Scientific Agenda, Kavli HUMAN Project; *Silver Professor* of Economics, Department of Economics; *Deputy Director*, Institute for the Interdisciplinary Study of Decision Making, New York University

LYNN GOLDSTEIN, *Chair,* Privacy & Security Advisory Council, Kavli HUMAN Project; *Fmr. Privacy General Counsel and Chief Privacy Officer*, JP Morgan Chase

GARY KING, *Albert J. Weatherhead III University Professor*, Department of Government, Harvard University

STEVEN E. KOONIN, *Associate Director*, Kavli HUMAN Project; *Director*, Center for Urban Science + Progress; *Associate Director*, Institute for the Interdisciplinary Study of Decision Making, New York University

JULIA LANE, *Professor*, Wagner School of Public Policy; *Provostial Fellow* in Innovation Analytics; *Professor of Practice*, Center for Urban Science + Progress, New York University

MARK LESLIE, *Lecturer in Management*, Stanford University; *Managing Director*, Leslie Ventures

KATHLEEN McGARRY, *Chair*, Study Frame Advisory Council, Kavli HUMAN Project; *Chair*, Department of Economics, University of California, Los Angeles

ARISTIDES A.N. PATRINOS, *Fmr. Deputy Director for Research*, Center for Urban Science + Progress, New York University

ALEX 'SANDY' PENTLAND, *Chair*, Measurement and Technology Advisory Council, Kavli HUMAN Project; *Toshiba Professor* of Media Arts and Sciences; *Director*, Media Lab Entrepreneurship Program, Massachusetts Institute of Technology

ELIZABETH A. PHELPS, *Julius Silver Professor* of Psychology and Neural Science, Department of Psychology; *Associate Director*, Institute for the Interdisciplinary Study of Decision Making, New York University

ROBERT J. SHILLER, *Sterling Professor* of Economics, Department of Economics, Yale University

MIYOUNG CHUN (*Ex Officio*), *Executive Vice President* of Science Programs, The Kavli Foundation

HANNAH M. BAYER (*Ex Officio*), *Chief Scientist*, Kavli HUMAN Project; *Research Associate Professor* of Decision Sciences, New York University

KAVLI HUMAN PROJECT STAFF

Staff Officers

SAMANTA SHAW, Chief Operating Officer

OKAN AZMAK, Chief Measurement and Technology Officer

CHRISTINE COWLES, Chief Study Frame Design Officer

LEWIS B. GROSWALD, Chief Education & Public Outreach Officer

Support Staff

MOHAMED Q. AMIN, Research Assistant

KATHLEEN R. FLAHERTY, Research Assistant

RIMMA ILYUMZHINOVA, Research Assistant

WENDEL B. SILVA, Research Assistant

ECHO WANG, Research Assistant

KYLE ANN STOKES, IISDM Program Administrator (through October 2015)

SAMEERAA PAHWA, IISDM Program Administrator (through May 2015)

MEASUREMENT AND TECHNOLOGY ADVISORY COUNCIL

ALEX 'SANDY' PENTLAND, *CHAIR*; *Toshiba Professor* of Media Arts and Sciences; *Director,* Media Lab Entrepreneurship Program, Massachusetts Institute of Technology

NADAV AHARONY, *Product Manager*, Android, Google

DENNIS A. AUSIELLO, *Jackson Distinguished Professor* of Clinical Medicine; *Director,* M.D./Ph.D. Program, Harvard Medical School & Massachusetts General Hospital

JEANNE BROOKS-GUNN, *Virginia and Leonard Marx Professor* of Child Development and Education, Teachers College and College of Physicians and Surgeons, Columbia University; *Co-director,* National Center for Children and Families; *Co-director,* Columbia University Institute for Child and Family Policy

JENNIFER KURKOSKI, *Research Scientist,* People Innovation Lab, Google

DAVID LAZER, *Professor* in Political Science, Computer and Information Science, Northeastern University; *Visiting Scholar,* Harvard University

KEVIN OCHSNER, *Professor and Director* of Graduate Studies, Department of Psychology, Columbia University

ROBERTO RIGOBON, *Society of Sloan Fellows Professor* of Management; *Professor* of Applied Economics, Massachusetts Institute of Technology

BENJAMIN SHILLER, *Assistant Professor* of Economics, Brandeis University

Program Officer

OKAN AZMAK, Chief Measurement and Technology Officer

PRIVACY AND SECURITY ADVISORY COUNCIL

LYNN GOLDSTEIN, *CHAIR*; *Fmr. Privacy General Counsel and Chief Privacy Officer*, JP Morgan Chase

JUSTIN BROOKMAN, *Director*, Consumer Privacy Project, Center for Democracy and Technology

JUSTIN CAPPOS, *Assistant Professor* of Computer Science and Engineering, New York University

MARTI L. DUNNE, *Associate Vice Provost* for Research Compliance and Administration, New York University

ROBERT M. GOERGE, *Senior Research Fellow,* Chapin Hall, University of Chicago

THOMAS HARDJONO, *Technical Lead & Executive Director,* The MIT Kerberos Consortium, Massachusetts Institute of Technology

JULES POLONETSKY, *Executive Director and Co-chair,* Future of Privacy Forum

HILARY WANDALL, *Associate Vice President, Compliance and Chief Privacy Officer*, Merck & Co., Inc.

MARCY WILDER, *Director,* Privacy and Information Management Practice, Hogan Lovells

MIRIAM H. WUGMEISTER, *Chair,* Global Privacy and Data Security Group, Morrison & Foerster

Program Officer

HANNAH BAYER, Chief Scientist

STUDY FRAME ADVISORY COUNCIL

KATHLEEN McGARRY, *CHAIR*; *Chair,* Department of Economics, University of California, Los Angeles

BJ CASEY, *Director,* Sackler Institute for Developmental Psychobiology; *Professor* of Developmental Psychobiology, Weill Medical College of Cornell University

BRIAN ELBEL, *Associate Professor* of Population Health and Health Policy, New York University School of Medicine

ARIE KAPTEYN, *Executive Director,* Dornsife Center for Economic and Social Research, University of Southern California

KENNETH M. LANGA, *Professor* of Medicine, University of Michigan

MATTHEW D. LIEBERMAN, *Professor* of Psychology, Psychiatry and Biobehavioral Sciences; *Director,* Social Cognitive Neuroscience Laboratory, University of California, Los Angeles

DEREK NEAL, *Professor,* Department of Economics, the Committee on Education, University of Chicago; *Research Associate,* National Bureau of Economic Research

PAUL THOMPSON, *Professor* of Neurology, Psychiatry, Radiology, Engineering & Ophthalmology; *Director,* NIH ENIGMA "Big Data" Center of Excellence; *Associate Dean* for Research, Keck USC School of Medicine; *Director,* USC Imaging Genetics Center, University of Southern California

Program Officer

CHRISTINE COWLES, Chief Study Frame Design Officer

SCIENTIFIC AGENDA ADVISORY COUNCIL

ANDREW CAPLIN, *CHAIR*; *Silver Professor* of Economics, Department of Economics, New York University; *Deputy Director,* Institute for the Interdisciplinary Study of Decision Making

DENNIS A. AUSIELLO, *Jackson Distinguished Professor* of Clinical Medicine; *Director,* M.D. /Ph.D. Program, Harvard Medical School & Massachusetts General Hospital

LAURA JEAN BIERUT, *Alumni Endowed Professor* of Psychiatry; *Co-Director,* Outpatient Psychiatry Clinic, Washington University School of Medicine

CLANCY BLAIR, *Professor*, Department of Applied Psychology, Steinhardt School of Culture, Education, and Human Development, New York University

JEANNE BROOKS-GUNN, *Virginia and Leonard Marx Professor* of Child Development and Education, Teachers College and College of Physicians and Surgeons, Columbia University; *Co-director,* National Center for Children and Families; *Co-director,* Columbia University Institute for Child and Family Policy

BJ CASEY, *Director,* Sackler Institute for Developmental Psychobiology; *Professor* of Developmental Psychobiology, Weill Medical College of Cornell University

DAVID CESARINI, *Assistant Professor* of Economics, Department of Economics, Center for Experimental Social Science, New York University

DAVID M. CUTLER, *Harvard College Professor, Otto Eckstein Professor* of Applied Economics, Harvard University

ADAM DREWNOWSKI, *Professor,* Epidemiology*; Director,* Nutritional Sciences Program, University of Washington

MASOUD GHANDEHARI, *Head,* Urban Observatory, Center for Urban Science + Progress; *Associate Professor*, Civil & Environmental Engineering, Polytechnic School of Engineering, New York University

PAMELA GIUSTINELLI, *Research Assistant Professor,* University of Michigan Institute for Social Research

EDWARD GLAESER, *Fred and Eleanor Glimp Professor* of Economics, Harvard University

CATHERINE HARTLEY, *Assistant Professor* of Psychology in Psychiatry, Sackler Institute for Developmental Psychobiology, Weill Medical College of Cornell University

ICHIRO KAWACHI, *John L. Loeb and Frances Lehman Loeb Professor* of Social Epidemiology; *Chair*, Department of Social and Behavioral Sciences, Harvard University

KENNETH M. LANGA, *Professor of Medicine,* University of Michigan

SCOTT LIPNICK, *Scientific Director*, Center for Assessment Technology and Continuous Health (CATCH); *Assistant* in Biomedical Physics, Department of Medicine, Massachusetts General Hospital; *Imaging and Data Specialist*, Stem Cell and Regenerative Biology Department, Harvard University

CHARLES F. MANSKI, *Board of Trustees Professor* in Economics, Northwestern University

KATHLEEN McGARRY, *Chair,* Department of Economics, University of California, Los Angeles

ARISTIDES A.N. PATRINOS, *Fmr. Deputy Director for Research*, Center for Urban Science + Progress, New York University

RUSSELL POLDRACK, *Professor,* Psychology; *Member,* Stanford Neurosciences Institute, Stanford University

C. CYBELE RAVER, *Vice Provost* for Academic, Research and Faculty Affairs, New York University

SCOTT SCHUH, *Director*, Consumer Payments Research Center; *Senior Economist and Policy Advisor*, Research Department, Federal Reserve Bank of Boston

WOLFRAM SCHULTZ, *Wellcome Principal Research Fellow, Professor* of Neuroscience, University of Cambridge

REGINA SULLIVAN, *Professor,* Department of Child and Adolescent Psychiatry, Child and Adolescent Psychiatry, New York University

CASS R. SUNSTEIN, *Robert Walmsley University Professor,* Harvard Law School

ROBERT M. TOWNSEND, *Elizabeth and James Killian Professor* of Economics, Department of Economics, Massachusetts Institute of Technology

TABLE OF CONTENTS

Appendicies not included in print version. Selected non-confidential appendicies available online at www.kavlihumanproject.org

TABLES & FIGURES

TABLES

**Confidential, Not Included*

FIGURES

ABSTRACT

While the United States alone has a long history of gathering data on its population, going back at least to the census of 1790, longitudinal studies of human populations conducted world-wide have almost uniformly been hand-crafted, small-scale enterprises aimed directly at answering a limited number of predefined questions. The U.S. Health and Retirement Study (HRS), perhaps the broadest and most influential longitudinal survey of all time, gathers surprisingly limited data on financial, social and health matters for older adults once every two years. It is aimed at understanding the impact of aging on social and physical well-being. But despite these limitations, the HRS has revolutionized our understanding of many aspects of aging. A review of longitudinal studies funded by the Kavli Foundation in 2013 concluded that the use of large-scale automated methods for capturing biological and digital data at a synoptic level could revolutionize nearly all aspects of the social and natural sciences, and that such a study was rapidly becoming financially feasible. That review proposed the development of a study methodology that would yield roughly 100x the data density of existing studies. It concluded that an ideally designed platform could achieve this data density at roughly one-fiftieth the cost per data point of existing methodologies.

This report presents this new study methodology, and proposes its use for initiating a highly detailed synoptic longitudinal study of 10,000 individuals in roughly 2,500 households over 20 years in New York City. All aspects of the study methodology are developed in detail from study frame design to data gathering technologies. It is proposed that these detailed data about individuals be complemented with a comprehensive archive of the environmental conditions and events currently being aggregated in New York City. It is then proposed that this combined longitudinal dataset be made accessible to the research community in order to facilitate previously unattainable advancements in medicine, biology, psychology, economics, sociology and anthropology while also fostering evidenced-based public policies. A series of White Papers by leading scholars included in the appendices of this report details the potential benefits of such a study.

Our proposed automated methodology would continuously collect an array of data that includes: genome, microbiome, metabolome, epigenetics, medical records, drug and chemical exposure testing, diet assessment, sleep behavior, physical activity levels and home air and noise quality. Personality traits, IQ, mental health status, long term and working memory, social network structure, communication partners and patterns, continuous location capture, educational history, employment data and detailed financial transaction records are also proposed amongst other data.

Were such a study to be initiated, it would provide for the development of a revolutionary platform for improving the human condition. Its development in New York City, with existing extensive city data sets, would provide the most detailed synoptic discovery data set on humanity ever gathered. This report provides a Preliminary Study Design for the development and deployment of this platform.

SUMMARY

Study Need

While there is a wealth of overlapping longitudinal studies being conducted today in the developed world, no methodology currently exists for gathering a truly synoptic dataset. The result of this limitation is that individual studies are designed to focus narrowly on a small set of measures that are hoped to yield insight into a single domain. In the 1970s, this was also true in many of the physical and genetic sciences. An astronomer hoping to gain insight into black holes manually searched the heavens for individual objects, data about which could be used to build a small database for the study of black holes. A geneticist hoping to explain psychosis might similarly examine single genes in laborious detail in an effort to build a local understanding of the roots of psychopathology. In the 1980s, many disciplines began to abandon this approach, however, in favor of more synoptic and automated methods. In 2000, the Sloan Digital Sky Survey (SDSS), for example, began the automated scanning of deep space in order to build a massive database that today profiles over 500 million celestial objects. The Human Genome Project (HGP) began the development of tools for automated large-scale genetic analysis in 1985 that now have yielded massive databases on the genetic structure of hundreds of organisms. Today, astronomers and geneticists achieve daily progress with these community resources that would have been unimaginable prior to their development.

Unfortunately, the synoptic study of human beings has proven more technologically difficult than either the study of the cosmos or of the genome. Human behavior, human biology, human development, human economic activity and human social interactions have all been studied successfully, but with labor-intensive tools that preclude the measurement scale required for a synoptic project like the SDSS or the HGP. Recently however, that has begun to change. Driven by corporate interests and rapidly maturing digital technologies, detailed measurements of nearly all aspects of human life and behavior are now being made in a piecemeal way by a host of large corporations. Were it possible to aggregate these existing technologies and develop a few key tools for supplementing them, a synoptic study of humanity could be possible today, at a surprisingly low cost. We estimate that the development and deployment of a synoptic study of humanity in an initial urban cohort of 10,000 could be achieved for an initial cost of approximately $10M/yr. Like the HGP, we believe that once the basic platform is developed, costs for future deployments could drop by at least a factor of 2 within 5 years.

Were such a database, or such databases, available to scholars, there is compelling evidence that a host of currently intractable medical, psychological, developmental, economic and sociological problems could be successfully engaged. Such a longitudinal dataset would, for example, definitively identify the many interlocking causes of our obesity epidemic. It would finally allow us to determine the critical factors in successful childhood development. It would define when and why people suffer from Alzheimer's disease by uniting data about environmental exposure to pathogens, diet, behavior, medical care and genetics in a single longitudinal database. It would identify the roots of psychopathology, financial success and educational attainment at unprecedented depth, all at a very low initial cost.

Management Plan

To achieve these goals we propose the development of a study we refer to as the Kavli HUMAN Project (**H**uman **U**nderstanding through **M**easurement and **AN**alysis) (KHP). We propose building the study around a variant of the Department of Energy's "Stage-Gate" process for large-scale study management. In this document we specifically propose a five-stage process: Stage 0: Study Need and Feasibility; Stage 1: Preliminary Study Design; Stage 2/3: Complete Study Design; Stage 4: Construction; Stage 5: Deployment. The Kavli HUMAN group delivered a Study Need and Feasibility analysis to the Kavli Foundation Board in October of 2014. This current document constitutes the Preliminary Study Design required next by our management plan.

Conceptually, the study development and execution process is organized around five principal divisions:

- Measurement and Technology; (What we propose to measure and how)
- Study Frame Design; (How we propose to select, recruit and retain subject-participants)
- Privacy and Security; (Security for physical and digital data, as well as participant consent)
- Education and Public Outreach
- Scientific Agenda; (How data collection meets social and scientific needs)

Each division of the study is currently governed by an *Advisory Council* of 10-15 leading scholars, technologists and social activists with a lead individual who serves as chairperson. Each of the five boards provides support to five complementary in-house staff groups responsible for executing all aspects of that study division. Each internal division is lead by an outstanding practitioner who works with their own staff to accomplish study goals in concert with the relevant Advisory Council.

Overall management of the study is achieved by the Governing Board of the study, through the Study Director who provides day-to-day management of the project. All five Advisory Council Chairs, the Study Director and five to ten additional members sit on the Governing Board of the study. Finally, the Director manages the study in concert with a Chief Scientist and an Administrative Director who oversee scientific and managerial issues, respectively.

Measurement and Technology

The utility of the entire undertaking proposed here rests on the notion that a fairly complete synoptic profile of the same 10,000 individuals can be achieved over 20 years at a reasonably low cost. A critical feature of this plan is that it must be accomplished in a sufficiently non-invasive manner that subject-participants both remain in the study and remain largely uninfluenced in their daily behavior by the study.

Several key assets make this possible. First, the development of digital data technologies by the private sector means that nearly all of the digital data we would require can now be gathered non-invasively and at scale. Second, the rapidly falling cost and wide availability of technologies for biosample analysis make a huge range of physical sampling technologies feasible at scale. Together, contemporary digital and physical sampling technologies place a detailed characterization of individuals within reach, given minimal technological and platform development. Third, the recent development of large-scale Geographic Information Systems (GIS) makes it possible to combine the proposed datasets with very detailed characterizations of our subject-participant's environment at a micro-level. GIS-based municipal and academic data about everything from local weather to hyperspectral image-based assessments of environmental toxins at the resolution of a city block can be combined with the participant-level data to complete the proposed dataset.

GIS databases of this kind are being developed throughout the world today. Although the most complete such database is now being built in New York City, comparable databases are under construction in many areas and the number of such

databases is predicted to grow radically in the next 5 years. While this makes New York City a logical site for platform development, we expect that within three years a number of additional sites will likely be available. Finally, we note that the 10,000 participants are embedded in roughly 2,500 households. Because families and households are studied together, many of our measurements (for example genetic measurements) have much higher statistical power than would be expected in a sample of 10,000 independent individuals. This sampling structure also assures that developing children are studied within the matrix of their families.

Demographics and Home Environment

The proposed measurement systems begin with standard demographic data about each participant of the kind gathered in a traditional study. This is supplemented by easily achieved measurements of the home environment, both derived from publically available databases and from a simple survey made by study personnel at the time of participant recruitment. NYC public datasets and GIS databases provide additional neighborhood-level data at a spatial precision of a single building lot and a higher than daily temporal precision.

Biological/Medical Samples

At the time of enrollment, and at roughly 3-year intervals thereafter, biological and physical data is gathered on each participant. We note that this is the most invasive (and expensive) part of the proposed study, and these measurements take place over the course of half a day. During this period we obtain a basic medical exam; a blood sample for genetics and blood chemistry; a urine sample for toxicology; a saliva sample for oral microbiome, genetics (in children) and salivary hormone analysis; a hair sample for toxicology and a stool sample for gut microbiome.

These biological samples are supplemented with medical data from the participant's electronic medical record, to which the study gains continuous electronic access. We also gain ongoing electronic access to all medical-diagnostic codes filed with the state and state prescription records.

Educational-Occupational

In addition to gathering traditional educational data, we propose to gain ongoing electronic access to Department of Education data on all young subjects and on the schools attended by our subjects. Similar data on college performance, when applicable, will be gathered from commercial databases. Traditional occupational data will also be gathered in an ongoing manner.

Financial

A detailed financial profile will be gathered at enrollment but more importantly, ongoing financial data will be gathered automatically at a daily level. Automatic web-based systems will allow data at the transaction level to be acquired for the vast majority of our participant's economic activity.

Smartphone

Perhaps the single most important data collection tool employed by the study will be each participant's smartphone. Smartphones allow us to gather a host of data about the participants including: continuous location data, activity data, and social network data from email, telephone and texting. The smartphones, and an accompanying Bluetooth beacon technology we are developing, also provide information on how and when family members – including children – interact. The smartphones also allow us to invite participants to complete very brief psychological and social questionnaires offered on a weekly basis. Aggregating data over many years these questionnaires provide an unparalleled data stream – the details of which are described in the body of this report.

Summary

The sections above capture many of the data streams that the study can gather. The detailed structure of our measurement process, how acceptability of these measurements to participants is assessed and maintained, and a host of additional measures, are discussed in the body of this report.

Study Frame Design

In order for the study to yield a meaningful characterization of the society in which it is conducted, it is essential that it be based on a randomized sampling strategy. Like the U.S. Health and Retirement Survey and other successful large-scale longitudinal studies, the proposed study must yield a quantitatively tractable cross-section of the study population. For the purposes of our initial study we propose here a cross-section of the city of New York, including all five boroughs. Restricting the initial study to New York City provides several advantages: it allows a sample of only 10,000 individuals to provide significant statistical power because it places them within a limited set of environments, it makes participant recruitment and tracking geographically tractable at low cost, and it allows us to leverage the data infrastructure of New York which is more advanced than any other urban environment at this time.

In order to maximize the power of the measurements we make on these participants we propose to study entire households. Our proposal is to begin by randomly selecting, using a statistical model of the city, roughly 2,500 "seed" individuals. The randomly selected seeds identify roughly 2,500 households – each of which becomes a participant family. Seeds are selected until a total study population of 10,000 is reached. Our analysis indicates that successful recruitment of 10,000 participants can be accomplished in 30-36 months at the proposed cost-level.

The sampling strategy is discussed in detail in this document. It rests on a flat sample of households anchored to a geographic study frame at the single dwelling level. A statistical oversample of three groups, young children, pre-teens and elders, is specified. Details of our sampling frame and model design are also discussed in this document.

Like many areas in the United States, New York is a multi-lingual city. We propose conducting the study in both English and Spanish. This allows us to reach better than 88% of New York City households at the high school language level. New York City also includes a significant group of Chinese and Russian speakers (an additional 6% of the population) as well as a host of other languages. We propose to develop a "Simple English" version of the study to increase the probability of reaching members of these difficult to reach populations.

Our proposed recruitment procedure is based on a modification of the standard Dillman method. This involves a recruiter making multiple incentivized contacts with a candidate household prior to recruitment. Our proposed method is a lengthened and more incentivized version of the technique employed by the Health and Retirement Survey that achieves roughly 85% recruitment rates in their population. The recruitment teams will be fielded in-house rather than being subcontracted to maximize our ability to develop and control the study platform. Our goal is to achieve a 25% recruitment rate in our population with our in-house teams.

We propose to employ three types of incentives: money and items that have a real monetary value, items and activities that build community good will and giving participants personal and population-level feedback and data. The literature shows that providing even a small monetary incentive (rather than none at all) creates an increase in response rates. Results from other longitudinal studies suggest that incentives other than money engender goodwill and feelings of trust with respondents, as people who feel part of something are more likely to continue participation, and over time the financial cost of these items is often less than would be spent for monetary incentives. Existing studies suggest that using these incentives should enable us to maintain a roughly 95% annual subject retention

rate after initial subject attritions. Treatment effects engendered by our incentive structure are discussed in the complete document.

Privacy and Security

The goal of the Kavli HUMAN Project is to create a resource for scholars that can be used to address a wide range of basic research and policy questions. However, this comprehensive data set will contain a wide range of sensitive material, including personally identifiable information, so access must be balanced against the need to protect the security of the data and the privacy of the participants. Successful management of the rich store of electronic data and biological samples is critical to the success of the project and thus, we have invested substantial resources in the design and implementation of this aspect of the study.

Our intention is to provide the highest possible degree of privacy and security to our subjects. To do this we propose to implement the NIST SP 800-53 security framework. This is a superset of the regulations required by the HIPAA and FERPA acts. It also details security protocols, data handling and data accounting procedures. In essence, our security plan rests upon a dual firewall system. Raw data resides inside a super-secure inner firewall accessible only to system operators. Data inside the inner firewall is partitioned into roughly six secure independent subdomains. Physical and digital access to data inside the inner firewall is only available after two-factor biometric identification of a sysop. Access to the data can, by hardware design, only be accomplished inside the physically secure data facility after two-factor identification of operators at the physical level. All system-wide operations require the consent of two system operators who have been physically and digitally two-factor identified.

In order to provide scholars with easy access to the dataset, we propose to provide a mechanism for scholars to request an anonymized subset of the data we refer to as a *datamart*. Datamarts are the object with which scholars interact, and they reside inside the outer firewall of the KHP. Datamarts, by design, do not contain personally identifying data. Scholars perform their analyses on datamarts to extract causal relationships, test hypotheses or to engage in data discovery operations. The product of these analyses can be exported outside the facility firewalls for publication or further analysis. All datamarts are destroyed after the study for which they were requested is complete. In addition to these secure data-types, fully anonymous statistical profiles of the database will be made available on the web.

In order to insure that none of the data gathered and stored can be subpoenaed, the data will be covered by an NIH Certificate of Confidentiality (NIH CoC). The NIH CoC insures that all of the data resident in the data storage facility is private and confidential, even in the face of a government or other third party request.

To ensure data survivability and recoverability, the entire database will be routinely encrypted using the AES 256-bit technology. The database will then be fragmented to non-recoverability and stored off-site in a 10-partition/10-location fragmentary backup via point-to-point transfers to remote backup servers. Note that no element of the database ever leaves the facility without these two stages of encryption: 256 bit AES encryption followed by encrypted fragmentation.

Finally, an external "red team," tasked with breaking down our security protocols, will continuously test the security of our system. This external group, discussed in the full document, is incentivized to identify weaknesses in our security system by placing it under continuous attack.

Education and Public Outreach

Education and Public Outreach (EPO) is a critical component of the proposed study, especially given the novel nature of the proposed study at this time. But for the KHP to reach its potential, open lines of communication must be fostered with all of the project's stakeholders – scholars, potential

participants and potential funders in both the foundation and government communities.

Probably the greatest challenge to the study is to ensure that potential participants understand the importance of the study and feel confident about the study's ability to protect and respect them. Many stakeholders initially express great concern over the highly detailed nature of the data we propose to gather and many scholars express concern that it will be difficult to identify participants willing to undergo this level of scrutiny. One of the central missions of the EPO team is to make sure that subjects understand that nearly all of the data aggregated by the study is pre-existing. One example of such a data type currently collected by existing technology is geo-location. Many potential subjects do not realize that corporate actors continuously track all cell phones with an accuracy of meters. One goal of the EPO team is to make sure that potential subjects view the study as aggregators of existing data rather than as creators of novel and worrisome datasets.

Perhaps even more important is that the study be seen as a public actor working for the good of the community. Potential subjects often express concern and discomfort with corporate data gathering. To eliminate this concern the EPO team will have to engage community leaders, and socially visible figures to insure that the goals of the project are well understood by potential participants.

The EPO team also must engage a number of other communities ranging from scholars to policy makers, and each will require an individualized message. Specifics of the task faced by the educational and public outreach unit are provided in this document in significant detail. Media plans for untoward events are also addressed.

Scientific Agenda

The proposed dataset is fundamentally an enterprise of *discovery* science. The goal of building the database is to provide a general-purpose tool, like the SDSS map of the cosmos, which can be used to understand almost any aspect of human behavior and biology. But it is equally important to demonstrate both *how* a database of this kind could be used and *what* a database of this kind could contribute. By developing 'use cases,' the builders of the database can develop a better understanding of what exact form the database should take. Perhaps just as importantly, powerful use case examples can reassure funding agencies that databases of this kind will quickly yield powerful translational insights.

For that reason the Scientific Agenda group is tasked with identifying leading scholars and asking then to consider how a database of this kind could be used to revolutionize their fields of study. Scholars report their conclusions in prospective analyses we refer to as "White Papers." This document includes information on over a dozen such White Papers from a range of disciplines. For example, a leading team of clinical scientists from Harvard University explains how such a database could provide the critical tool for developing a new kind of "Real Time Medicine" that could be deployed in any population of patients. Another team from the University of Michigan and Harvard University explain how such a dataset could finally explain how and why cognitive decline proceeds at such different paces in different individuals. A team from Washington University and New York University describe how this dataset could resolve fundamental questions about who is at risk for different forms of substance abuse. A team from the University of Washington and Harvard University explain how this dataset could be used to get to the heart of the problem of obesity. A team from the University of Michigan and Northwestern University explain how these data could resolve important questions about what kinds of educational options are best suited for different kinds of students. Other papers explore the importance of "field studies" of neuroscience, the need for longitudinal studies of cognitive function, the prevalence of child abuse and the design of cities and the nature of psychological development in adolescents.

Ongoing work by the Scientific Agenda group seeks to expand the domains of inquiry we examine so as to test and retest the utility of the current study

design. Details of these high impact uses of the proposed database can be found in the Scientific Agenda section of this document. Complete copies of many of these White Papers can be found in Appendix K.

Appendices

The complete document also includes a detailed risk analysis and risk management plans for the proposed study. This risk analysis examines and evaluates all possible risks to the study and its participants. Risk mediation plans are proposed for all high risk items.

A detailed acquisition plan is also provided. This takes the form of a preliminary year-by-year budget for the proposed study. It identifies all costs and relates these costs to proposed funding sources. The acquisition plan precedes a more detailed financial model that will be developed for the final study design.

Specific reports from each of the Advisory Councils are attached. These documents reflect preliminary meetings by the Councils and served as the basis for developing the detailed documents in each of the five study domains.

Some of the measurements proposed in the Measurement section rely on novel technologies developed in house and a detailed assessment of existing wearable technologies. A report on the novel Bluetooth technologies, which are of particular importance for the study of younger children is the first of these sections. The second provides a detailed study-specific assessment of the utility of existing wearable technologies for passive data gathering.

A report by the leading privacy and security law firm Hogan-Lovells, commissioned by the study, is included to provide an external assessment of the privacy and security risks faced by the study.

A proposal from the Synack Corporation for the development of "red-team" IT capabilities to ensure ongoing testing of the database is appended next.

An overall communication plan commissioned from the leading market research firm Berlin-Rosen is appended to provide an external view of the risks and challenges faced by the study in its external communications.

A summary of the acceptability of intensive data gathering of this kind to potential subjects is presented. This independently commissioned study was funded by the NIH in order to assess the willingness of subjects to participate in the NIH Precision Medicine Initiative. It was conducted by the market research firm GfK and provides an excellent overview of subject acceptability issues for English and Spanish speaking US populations.

Biographies for the members of the KHP governance team project are also included to provide a view of the prestigious scholars who advise on decisions about study design and implementation.

Finally, the Appendices conclude with a set of representative Kavli HUMAN Project White Papers developed by the Scientific Agenda Group.

Conclusion

Based on this detailed analysis the study group has drawn the conclusion that it is now possible to develop a platform for large-scale data gathering in a representative human population at an acceptably low cost. Like other large-scale discovery projects conducted over the last three decades, this platform could revolutionize our understanding of its study target: human beings behaving in natural environments. Our analysis suggests that although this could be one of the highest impact discovery dataset projects ever undertaken, it would also be one of the cheapest. Indeed, were the platform to be developed, it would likely reduce the cost of data gathering about humans by a factor of 20-50 times. The existence of such a resource would also eliminate the need for the many partially

overlapping studies conducted today, suggesting very high overall cost savings.

More importantly, the study would advance a completely novel depth of understanding, of the kind accomplished by the Human Genome Project, in the domain of human brain, behavior, biology and society. While it seems clear today that this deeper understanding would yield many societal and scholarly benefits of the kind documented in our White Papers, it is impossible to say with any certainty where the revolution engendered by such a study would stop. Just as it would have been difficult in 1990 to imagine the genetic revolution of today brought on by the Human Genome Project, it is very difficult to say where the HUMAN Project will lead us. As Prof. Steven Koonin, a former Provost of CalTech and former Under Secretary of Energy for science recently put it:

> *"Great science always comes from new instrumentation. When Galileo first turned the telescope on the heavens it opened up a great vista for understanding our place in the universe. When van Leeuwenhoek first looked at a cell through the microscope it opened up a whole vista of understanding biological systems. I think that the Kavli HUMAN Project, with the technologies, the data analysis, and the understanding we have of humans now, has the potential to do the same sort of thing for individuals and the society that is made up of them."*

1

STUDY NEED

1. Introduction

Sociologists, political scientists, economists, psychologists, neurobiologists and other behavioral scientists have advanced rich theories of individual behavior and group interactions. Our ability to test these theories, and ultimately our ability to understand human behavior, rests on datasets gathered by these scientists, by governments and by corporations. A review of existing datasets, these catalogues of human behavior that guide scholarship and policy in the United States and throughout the world, reveals both their power and their limitations. The power of these datasets is that they provide detailed catalogs of genetic data or data about finances, or data about cognitive function. Their great limitation is that no single study examines all of these aspects of human behavior, biology and environment in a single group of subjects.

Just as geneticists working in the 1970s consisted of camps of scholars with deep expertise on isolated genetic systems, scholars of human behavior remain largely Balkanized into groups with deep, but local, expertise in specific aspects of human behavior. In the 1970s the US scientific community responded to the challenge of a fragmented genetic community by proposing and executing the Human Genome Project. Rather than a piecewise approach to understanding genetics, a group of visionary scientists proposed a complete catalog of the human genome – a proposal to transform genetics from a series of isolated fiefdoms into a global synthetic field. In proposing that vision, the human geneticists borrowed the language and authority of large-scale physical science. Since before World War II, physical scientists had responded to large-scale challenges with large-scale consortia that have ranged from the Manhattan Project to the Hubble Space Telescope. What they produced - the consensus sequence of the human genome, a new class of bioinformatics, the tools for rapidly sequencing the genomes of other species, even the ability to fully sequence the genomes of individuals at low cost – have revolutionized the biological sciences.

Here we raise the possibility that it is now technologically feasible to undertake a large-scale measurement of human behavior that spans the sciences from sociology to neuroscience. Our goal is to begin by capturing, for 10,000 people, in approximately 2,500 households, living in a major metropolitan area, literally all aspects of their human lifecycles. Beginning with genome and microbiome data, developmental psychological measurements, cognitive testing, educational tracking, personality analysis, economic development, social networks, health and retirement, to name just a few areas, we propose a deep and fundamental change in how we understand human behavior. Our goal is to develop the foundation for an International Observatory for Human Behavior. Just as the development of advanced optics and computers led to the development of national observatories for understanding the heavens a few decades ago, we believe that the time is right for turning our scientific lens toward the study of our most valuable asset: human behavior.

The power of a comprehensive approach of this kind can be seen in the Sloan Digital Sky Survey

(SDSS). Prior to the development of the SDSS, data collection was dominated by individual astrophysicists competing for limited access to telescopes, which they used to collect limited data sets that were primarily useful for addressing a specific question. The SDSS revolutionized the field by shifting to a new approach - a systematic examination of the sky. The most efficient instruments were used to build a comprehensive, publicly available dataset applicable to a broad range of questions in astrophysics. This resource has become invaluable to astrophysicists, supporting thousands of publications by researchers from around the world. We envision that a comprehensive survey of human behavior could similarly galvanize the social sciences.

2. Mission

Focused large-scale longitudinal studies have historically provided critical scientific insights. For example, the medical sciences were, in large measure, revolutionized by what has come to be known as "The Framingham Heart Study," begun in 1948. In that visionary project, the extremely detailed health (and behavioral) data of roughly 5,000 subjects was followed for what was initially imagined to be a 20-year period. At the time, the data gathered about the cardiac health of these subjects was at an unprecedented level of detail and the analysis of that data provided many of the fundamental insights about cardiac health that now serve as centerpieces of international healthcare. In the domain of social sciences, the power of this kind of approach was employed in 1992 when the National Institute of Aging created the U.S. Health and Retirement Study (HRS). This study behaviorally and financially profiles more than 20,000 Americans once every two years with a questionnaire – and recently with a request for genetic data. The results of those questionnaires are then made available to scholars for purposes ranging from assessments of US retirement and health care policy to the development of models of cognitive decline over the lifespan. And despite the fact that the HRS asks only a handful of questions of its participants every two years, it has served as the gold standard for understanding and characterizing human behavior in its domain. Indeed, studies of the kind pioneered by the HRS have now been conducted in dozens of countries.

What has never been attempted, however, is a large-scale and highly detailed measurement of human behavior – an effort to build a broad-scale set of behavioral measurements that rival the detail of the Framingham study's medical questions – but in domains ranging from the biomolecular to the sociological. We propose just such a study, initially in a limited set of 10,000 individuals making up 2,500 families in a major metropolitan area. Our goal would be, over a 20-year period, to make highly precise and detailed measurements of behavior and outcomes across all the important domains, not just those related to heart health. Such data would provide a database for the nearly complete characterization of behaviors of scientific and policy interest – a "human genome project" for behavior. What we believe makes this possible today are two huge advances in technology: the computer-based tools for aggregating and analyzing truly massive datasets and the development of new technologies, like smartphones and web portals, which allow researchers to follow subjects at an unprecedented level of detail as they move through all aspects of their daily lives.

A critical feature of human behavior is that the whole is more than the sum of its parts. Human behavior emerges as a complex system from interactions at many scales. Social and natural scientists across the many disciplines that study human behavior have worked hard to demonstrate this fact. They have accomplished that by working to mate existing datasets, by working to develop novel datasets and by seeking to expand outside their native disciplines.

Over the last decade there have been significant efforts to combine studies at different levels to derive a more holistic picture of human behavior: Social network studies have begun to make personality measurements. The HRS has begun genotyping. We take these as clear evidence that the need for synthetic measurement is pressing. Such

studies are beginning to form painstaking one-at-a-time linkages, just as studies of genetics began to build linkages in the 1970s. But what we propose here is a concerted effort to develop a large-scale database that links all of these domains in a single pool of subjects, initially in a single large-scale study to be conducted in New York City. Our goal, however, is for these focused measurements in New York to serve as the nucleus for a worldwide effort to understand and catalog human behavior. We stress from the outset that the goal of this proposed study is fantastically inclusive. Our intention is to bring together biologists, psychologists, economists, sociologists and anthropologists to develop a truly interdisciplinary catalog of human behavior that can play a transformative role in science and public policy.

2

PROJECT MANAGEMENT OVERVIEW

1. Introduction

An essential ingredient, perhaps *the* essential ingredient, in any large-scale project is the system selected for project management. In order to develop a system for project management and review the leadership of the Project, KHP began with a review of management systems employed at NASA, the NSF, the Department of Energy (DoE) and the NIH. Essentially two basic approaches to project management are employed in these agencies. In what is called the "Stage-Gate" model, projects undergo rigorous review at discrete stages that serve as stepping-stones. Each stage of review must be completed successfully before the next stage is undertaken. Typically, these stages include at the least:

- A "Need Statement" which succinctly describes the goals and value of the proposed project. At the DoE, review of this document is referred to as the "Critical Decision Zero" (CD0) review.
- A very detailed "Preliminary Design" which includes a complete analysis of every aspect of project design, project management and financial design as well as a detailed risk analysis. At the DoE, these documents typically span 200-400 pages in length. Reviews of these documents at the DoE are referred to as the Critical Decision One (CD1) review.
- A "Complete Initial Study Design" (CD2) that fully details every aspect of the project in sufficient detail for direct implementation.
- A "Final Study Design" (CD3)
- A "Fabrication Stage" (CD4)
- A "Project Initiation Stage" (CD5)

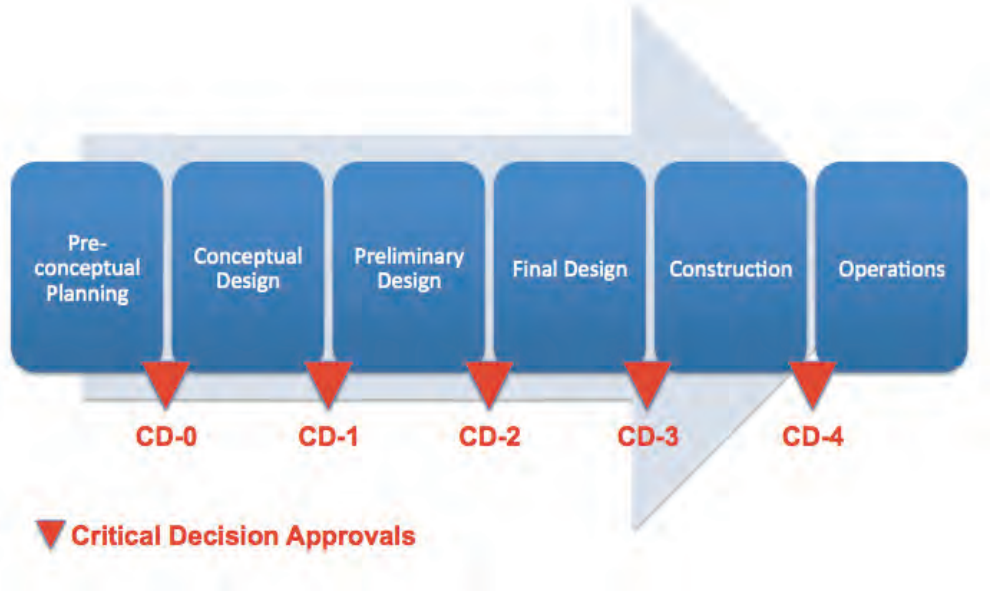


Figure 2-1: *The DoE Stage Gate Model*

An alternative approach employed with particular frequency at NIH is to request a 12-24 page project overview which combines elements of the "Need Statement" and the "Preliminary Design" which is the subject of a single detailed merit review. All subsequent stages in this project model are then managed internally without external review by the project team.

2. Management Model

The KHP leadership elected to develop a hybrid model very closely aligned with the DoE "Stage Gate" model. Our historical review of large-scale project success rates at NASA, NSF, DoE and NIH clearly indicated that project success (which we define as meeting the goals stated in the initial study need document, ideally on time and on budget) rates are much higher for "Stage Gate" managed projects with the absolute highest success rates being achieved by "Stage Gate" projects at DoE and NASA. Most DoE and NASA projects, however, involve the development of a large physical instrument, which places a higher emphasis on the development and assessment of blueprints and fabrication facilities. We thus elected to adopt the DoE model, collapsing the CD2 and CD3 stages into a single stage – as is occasionally done for projects of this type even at the DoE. Our staging process is thus:

- CD0: Study Need
  This 20-40 page document was tasked with assessing the *scientific value*, *cost* and *feasibility* of the proposed study.
  The actual Study Need Document was submitted to the Kavli Foundation for review in late 2014.

- CD1: Preliminary Study Design
  This approximately 500-page document was tasked with providing a complete overview of every aspect of the study. It was to provide
  - ☐ a management and governance structure for the study
  - ☐ a detailed measurement design for the study
  - ☐ a detailed study frame design for the study
  - ☐ a detailed privacy and security plan for the study
  - ☐ a detailed education and public outreach plan for the study
  - ☐ a very detailed assessment of the scientific impact of the study
  - ☐ a complete risk analysis for the study
  - ☐ a projected timeline for the study
  - ☐ a detailed financial overview of the study by year, and by budget category
  - ☐ a biographical section describing the capabilities of management team members

  This document you are currently reading serves as the submitted CD1 document for the study. It was delivered to the Kavli Foundation at the end of the third quarter of 2015.

- CD2/3: Final Study Design
  This roughly 1,000-page document will be tasked with providing a complete finalized design of all aspects of the study. It serves as the final study blueprint and must be reviewed and approved before the study can begin. It must provide sufficient detail on all aspects of the study for full implementation.
  CD2/3 is scheduled for completion at the end of the third quarter 2016

- CD4: System/Infrastructure Construction
  During this stage of the project a required infrastructure is completed, all staff required for study onset are hired and trained. The data warehouse, recruitment teams, outreach vans, office structures, every required aspect of the study is built. The CD4 stage completes with a successful test run of the study.
  CD4 is scheduled for completion at the end of the first quarter 2017

- CD5: Project Run
  Project Run is scheduled for the beginning of the second quarter of 2017.

3. Basic Administrative Structures

After a careful analysis of the tasks outlined in the CD0 document, the leadership of the KHP elected to develop the study around 5 principal domains:

1. Measurement and Technology
2. Study Frame
3. Privacy and Security
4. Education and Public Outreach
5. Scientific Agenda

Oversight of each of these domains is to be provided by a 10-15 person Technical Advisory Council recruited as volunteers from the very highest levels of the Academy, Government and Industry. Members of the 5 Advisory Councils should be individuals widely recognized to be of extraordinary accomplishment. Each Advisory Council is directed by a single Chairperson, who assumes principal responsibility for the quality of the study in their Council's domain.

At this time all 5 advisory councils have been staffed. Council members and council chairs interact regularly with KHP staff and meet occasionally as a full unit. Our target is for Councils to meet at least annually. Chairs interact at least weekly with KHP staff and guide staff in their regular interactions with all Council members.

Each of the 5 advisory councils serves specifically to guide 5 members of the KHP staff known as Officers. These Officers serve as division heads and report to their respective Advisory Council Chairs. Thus, to take one example, the Measurement and Technology Chair and the Measurement and Technology Officer serve as the essential dyad for all Measurement and Technology issues. Under extraordinary conditions, an advisory council chair may also serve as an officer. This may be particularly relevant with regard to the Scientific Agenda group, where academic credibility is critical in the Officer.

Two external entities provide oversight and guidance to the project. An Oversight and Review Council, initially staffed and maintained externally by the KHP's funding organizations, provides an independent critical overview of all aspects of the project. This group also makes recommendations for Critical Decisions to the Board of Directors at each project stage. The IT Red Team reports to the Chief Measurement Officer, but retains complete functional independence. Its mission is to identify privacy and security weaknesses in the KHP data infrastructure.

Central governance of the KHP is provided by an overall Board of Directors that has final approval over all aspects of the study. The Board of Directors is composed of the Chairs of each of the 5 advisory councils, The Study Director who oversees all aspects of the study on behalf of the Board, and 5-7 additional Board members.

The Study Director provides direct oversight over the activities of all Councils and is directly responsible to the Board for all aspects of the study. Administratively, he or she achieves this through two key staff members: The Chief Scientist and the Executive Director. The Chief Scientist is responsible for all 5 of the Officers who report directly to her. The Executive Director is responsible for all organizational and administrative functions. All administrative staff report to her, and she in turn reports to the Director.

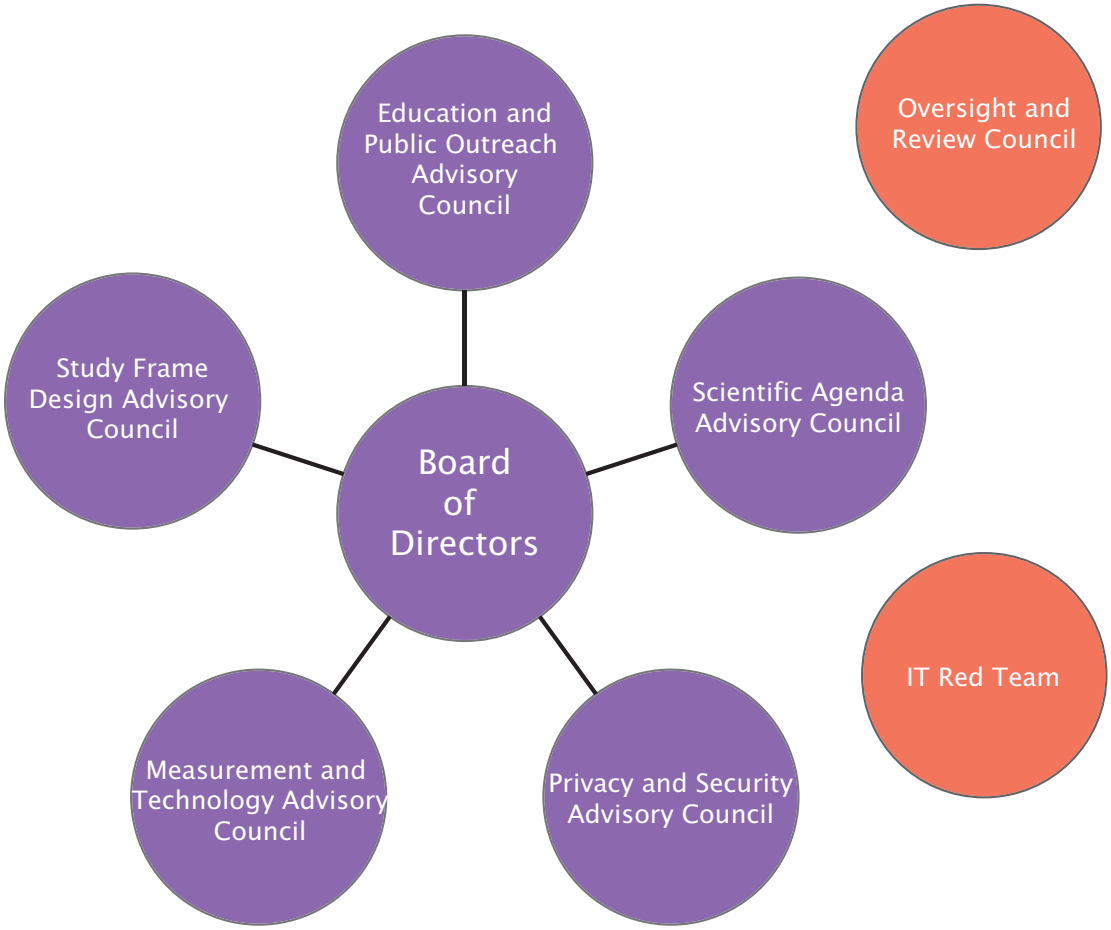


Figure 2-2: *Administrative Structure* shows these 5 divisions of the Council Structure. Also shown are the external review group and the fully independent "red team" for assessing the effectiveness of the Privacy and Security Council.

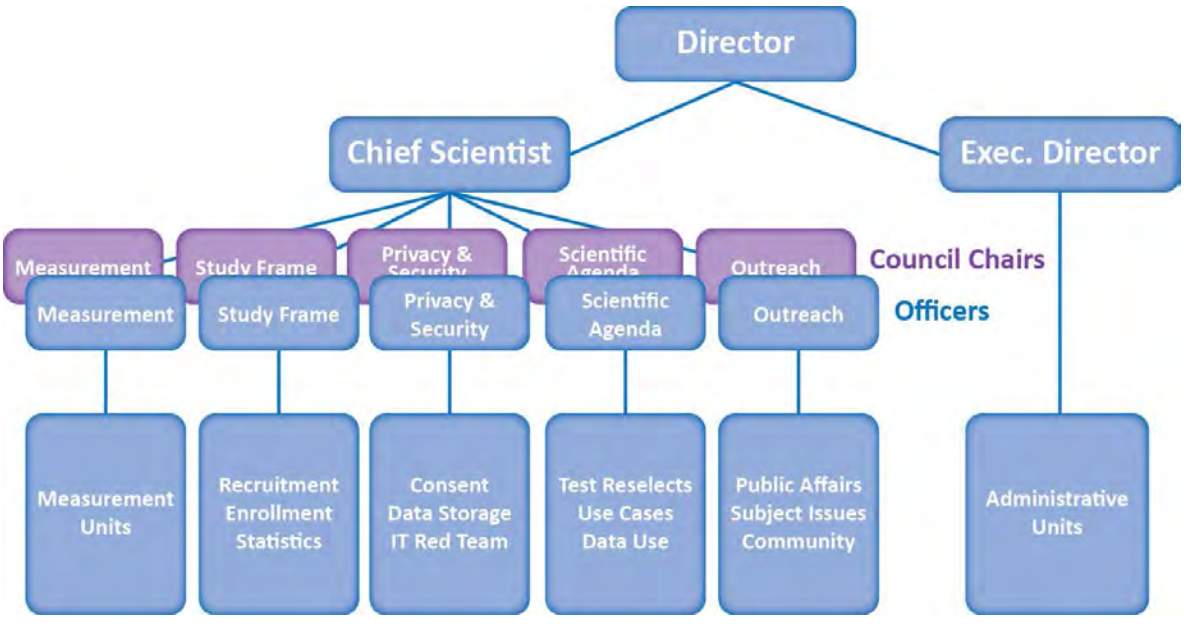


Figure 2-3: *Institutional Organization Chart*

3

MEASUREMENT AND TECHNOLOGY

Summary

Until now, large-scale longitudinal studies have generally been focused on specific domains of inquiry or subsets of the population. These studies typically provide detailed catalogs, in siloes, of genetics, health records or data about finances. At times, such studies generate integrated data about health and finances, but to our knowledge no study to date has examined the complete dynamic pattern of human behavior, biology and environment across the lifespan in a single group of subjects. The Kavli HUMAN Project (KHP) plans to address this gap by gathering information across numerous domains to provide a 360^0 view of the study participants. It is this synoptic view of 10,000 participants that makes the KHP unique. It is also true, however, that it would be impractical to aspire to measure literally every aspect of the human experience. Some classes of measurements are simply too expensive to make at scale. Others would be unacceptably invasive for subjects. Yet others offer too little value for the more modest expenditures that they would require. It is thus of critical importance that the study design process be engineered so as to correctly identify the measurements of greatest utility to the scientific community, given reasonable limitations on budget and subject acceptability.

With these issues in mind, this document provides:

1) A brief overview of the scope of data collection envisioned for the study
2) A discussion of the study design process for selecting measurements for inclusion
3) A proposal that the study cohort include three nested cohorts, each of which is the subject of slightly differing measurement densities
4) An overview of the data collection and data ingestion process including a discussion of the KHP Application Programming Interface (API)
5) A detailed overview of the data types planned for collection
6) A spreadsheet of every proposed measurement, including information on measurement frequency

1. Overview

The goal of the KHP is to provide as nearly synoptic an overview of human life as possible, initially profiling residents of the city of New York. To this end, we begin by identifying 3 main classes of data which will form the backbone of the KHP: 1) Data about the environment in which our participants live and work, 2) Physical, Biological, Psychological and Life Experience Data drawn directly from our participants, and 3) Digital data gathered at a subject- specific level from a number of interrelated sources.

1.1 The Physical Environment

The data about the physical environment in which our subjects live and work, data about the city of New York, must be gathered and curated with great depth and precision. Fortunately, the KHP will have permanent access to the ongoing database of the Center for Urban Science and Progress, or CUSP.

CUSP is a joint venture by New York City and New York University (NYU) that aims to provide a complete physical picture of the city. By agreement with New York City, CUSP has direct access to municipal data flows of nearly all types. This includes data on the K-12 education system, transportation, energy use, traffic and moving violations, law enforcement and legal system, among other domains. It is important to note that the breadth of the CUSP dataset grows continuously year-by-year. For example, CUSP is now exploring the use of real-time hyper-spectral imaging to track pollutant densities across the urban environment and to incorporate these chemical densities into its database. In essence, the CUSP database constitutes one of the most ambitious Geographic Information Systems (GIS) ever aggregated: a block-by-block, moment-by-moment, searchable record of nearly every aspect of the New York City landscape. (While the CUSP database really is unique at this time, a number of other cities are rapidly aggregating similar databases. Chicago, for example, should have completed a similar project in just a few years. Related projects are underway outside the United States as well, notably in London and Singapore. This is good news, because it means that studies like the KHP should become possible in many locales very soon.) CUSP data sets will be complemented by Census data to gain additional insights into the immediate demographic and socio-economic environment in which our subjects live and work. Finally, to gather information on subject's immediate environment, a detailed analysis of the home will also be gathered, including such critical data on the layout and size of the living spaces, presence and structure of the kitchen, etc.

1.2 Physical, Biological, Psychological and Life Experience Data

While this detailed portrait of the city is critical, data about our participants is the principal focus of the study and the unique focus of the KHP-team's data collection efforts. We propose to gather a range of physical and biological data to develop a portrait of our subjects. We begin with blood samples for health and nutrition profiling at intake. These blood samples will be used to gather data on diabetic status (glucose levels and glycohemoglobin), heavy metal levels, cholesterol levels, liver function, kidney function and a complete blood count. Blood samples will also be used for detailed genetic analysis. We propose to complete, at a minimum, whole exome sequencing for each participant and whole gut microbiome sequencing. Epigenetic, metabolomic, limited proteonomic and exposome analyses will also be conducted. Drug metabolites (for both legal and illegal, controlled and uncontrolled drugs) will be assessed. Urine samples and hair samples will provide detailed data on both present consumption and past consumption. (Hair analysis can, for example, determine tobacco consumption on a month-by-month basis years into the past.) Basic biometric data like blood pressure, height and weight will also be gathered during intake. All measurement values and test results will be converted to a standardized electronic format and stored in digital form. (We note that biological samples will also be stored for future use to leverage expected cost reduction for existing analyses, and in anticipation of new analyses that do not yet exist.) A more detailed description of biological specimens is provided in the later sections of this document.

In addition, at intake, a detailed set of psychological tests including IQ and a psychiatric evaluation will be performed. Finally, a detailed medical history, a detailed educational history, a detailed work history, a detailed financial history and a standard family history will also be gathered at intake.

1.3 Digital Data

During the course of the study, both structured and unstructured digital data will be gathered on our participants using a variety of techniques. Participants will be asked to direct copies of their existing data flows in the digital world to KHP databases. These flows include medical, educational and financial data. KHP will gain access to electronic medical records (NYU and a consortium of other New York City hospital systems is currently building a city-wide electronic medical record system. We anticipate that this will cover about 70%

of our subjects), and to state-gathered synoptic ICD-9 diagnostic code databases for each individual, along with access to insurance data. KHP will gather financial data, such as bank account and credit card transactions, through applications such as Mint.com and through 3rd party data sources, such as Yodlee, which aggregate this information. Lastly, KHP will access individual-level educational data in the NYC Department of Education that are available for research. These can also be supplemented with purchased data sets aggregated for us by companies like Experian (gathered with the consent of our participants), when that proves more cost efficient than more direct methods.

The homes of our participants will be equipped with a small base-station, which will report at-home measurements, such as environmental luminance, air quality and, most importantly, Bluetooth data from the subjects to the KHP data center. This base station is, in essence, a programmed smartphone; and it provides a secure data channel to the KHP datacenter at low cost. Each participant over the age of 10 who possesses a smartphone will have that smartphone fitted with the 'KHP app'. The KHP app is a simple self-monitoring program that serves as a smart gateway for several forms of data collection. The smartphone app collects geolocation of participants at regular intervals, allows cognitive tests of many kinds to be presented to the users, and provides a day-by-day, hour-by-hour link between participants and the Project. Critically, the KHP app is a very sophisticated self-monitoring device that alerts the KHP when there are data collection problems, but never bothers subjects or requires their help in its own maintenance. The smartphone app will be supported on a wide range of platforms, both Android-based and iOS-based operating systems. The app allows the KHP to set daily, weekly and monthly time budgets for interacting with the subjects. All tests, surveys and questionnaires delivered by the app are brief (always < 3 minutes), polite and game-ified so that they are a pleasure to complete. The app is smart enough that it can learn when subjects prefer to perform tests and surveys and respect the sleep, social and activity patterns of the subjects.

Subjects over the age of 10 who do not possess smartphones will be fitted with smartphones by the KHP. Critically, these smartphones will be delivered to subjects with voice and data contracts paid for by the project that make continued possession of the smartphone by the participant highly desirable. (Our goal here is to minimize the incidence of participants selling their smartphones.)

Subjects under the age of 10 will instead be outfitted with Bluetooth beacons with a 6-12 month lifetime. These allow the base stations and smartphones to determine the location of young children in the home, both with regard to home geography and distance from the adults. We are currently studying the feasibility of a disposable Bluetooth device that can track and store distance to all other Bluetooth devices (including smartphones) and can download that data to the home base station at regular intervals. Such a device would also include a multi-axis accelerometer for activity tracking. These disposable smart-beacons would serve as the central data collection tool for our youngest participants. We are currently exploring the possibility that these disposable 'smartbeacons' could be integrated permanently into the underwear of "KHP Kids".

Finally, we are exploring the utility of equipping adult subjects with activity trackers in addition to smartphones. Digital accelerometers embedded in smartphones do a good job of gathering activity data as long as the smartphones remain in a participant's pocket. But once a smartphone is placed on a table or in a handbag, this capability is lost. Activity trackers provide a supplementary capability, in this regard, but our initial test results with them have been mixed. The key problem is that activity trackers tend to be discarded, even by highly motivated subjects, after a period of several weeks. Our experience suggests that only autonomous, attractive waterproof trackers with very long battery life (> 6 months) may be appropriate for use in the KHP. Such devices are now coming to market and we plan to assess them in the next few months.

The paragraphs above provide a very brief overview of the scope of data collection for the KHP. In the sections that follow we provide more detail on both

the scope and mechanisms for data assessment and collection.

2. Measurement Selection Process

Over the last century, survey-based methods have been increasingly used to describe many features of 'the human experience' in a range of communities and settings. The goal of all these studies is to capture data that describes some limited aspect of how humans and their biology interact with environment.

Figure 3-1 captures the set of all possible study ranges in a graphical format. On the horizontal axis, the figure plots the age of the studied participant(s), ranging from birth to death. The left vertical axis, in contrast, plots the scale of the objects of measurement. At the very bottom of the graph, one can see measurements inside the nucleus of a cell (for example measurements of the genome), and at the very top one can see measurements at the scale of communities ranging in size from a neighborhood to a more global scale. The axis on the right relates each of these scales to the academic fields most traditionally associated with those measurements, ranging from Biology to Sociology and Anthropology.

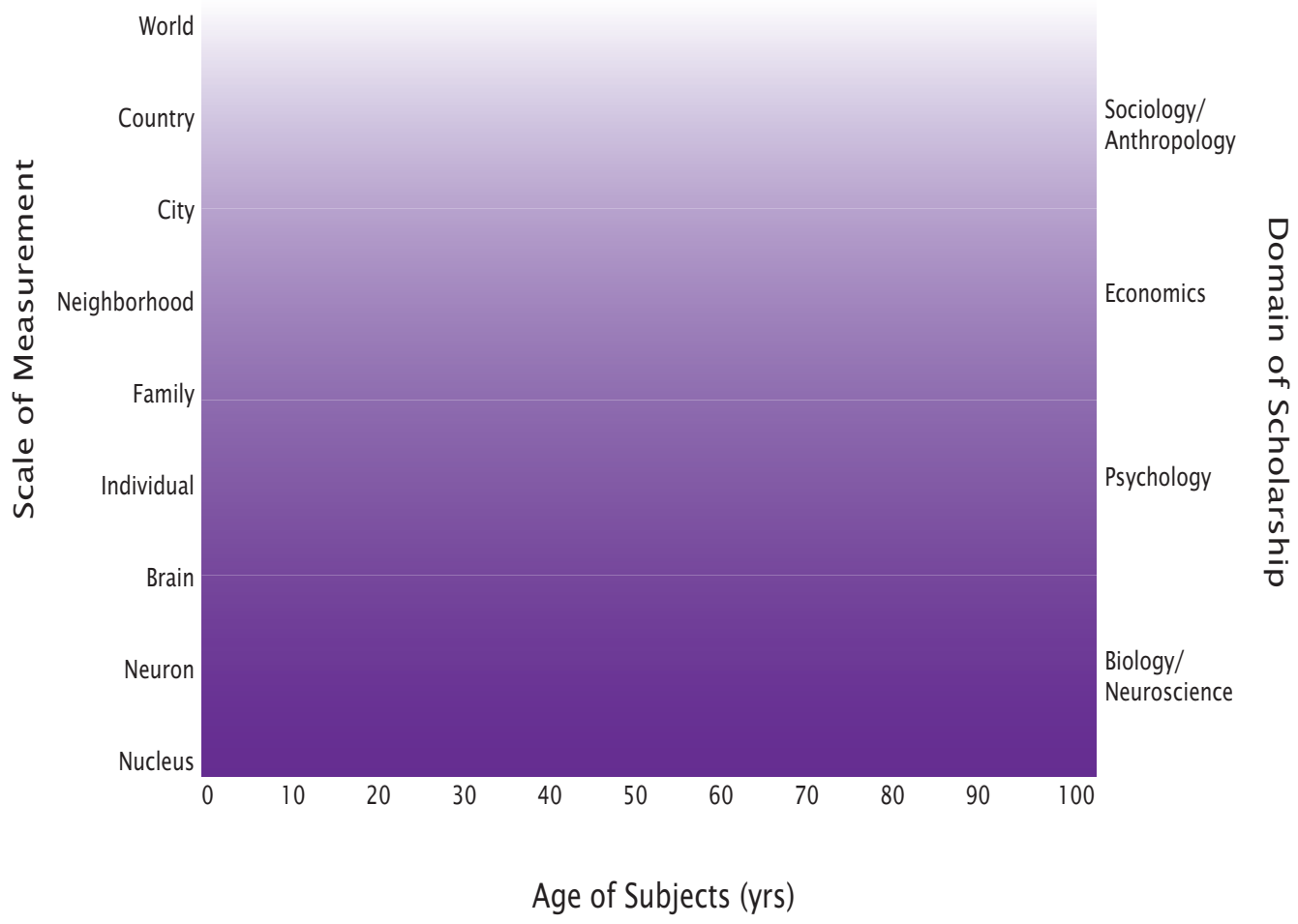


Figure 3-1: *The Space of Human Behavior*

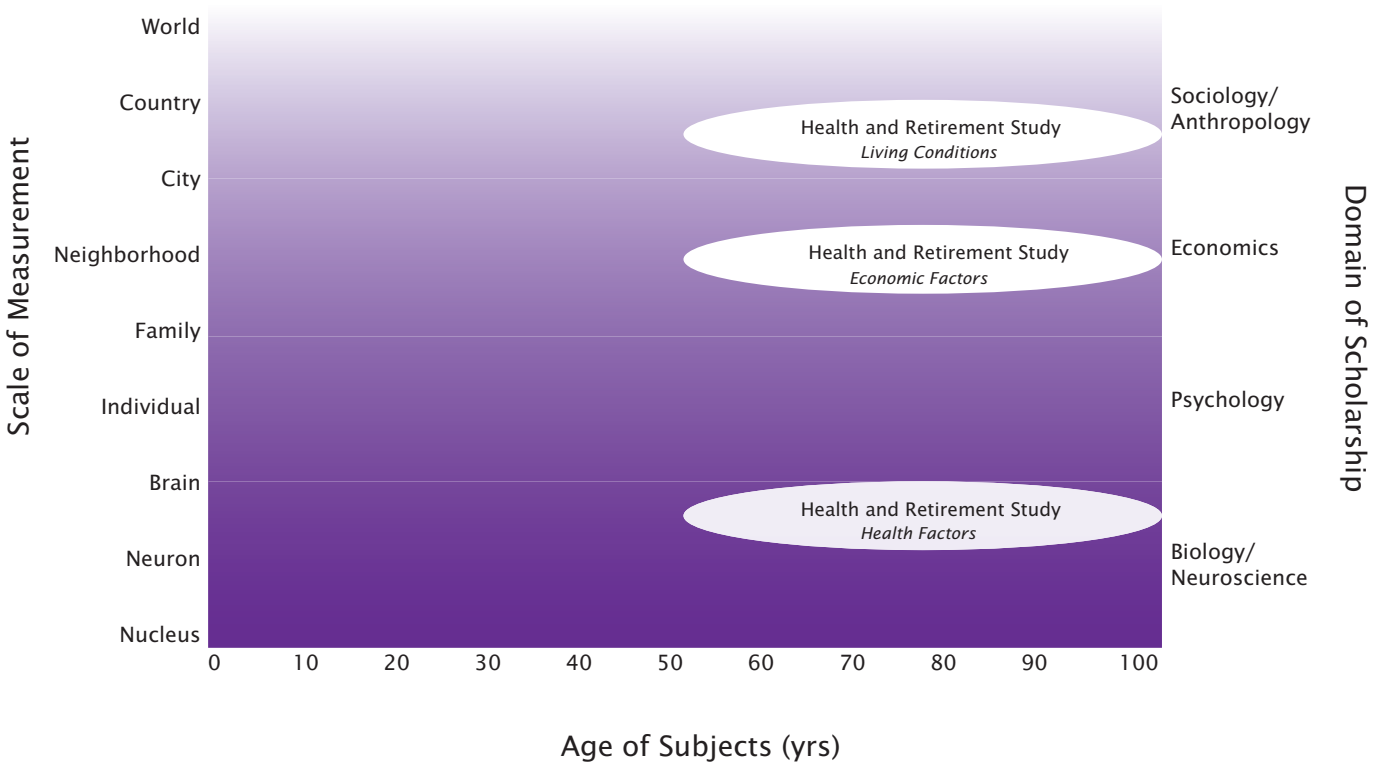


Figure 3-2: *The United States Health and Retirement Survey*

With this mapping in hand, it may be instructive to consider existing large-scale research projects. To do that, we turn first to what we consider to be the gold-standard for large-scale studies in the United States: The U.S. Health and Retirement Survey (HRS) conducted by the Institute for Social Research and funded by the U.S. National Institute for Aging. The HRS was designed decades ago to longitudinally profile U.S. residents above the age of 50. The HRS can be seen in our 'dataspace' in Figure 3-2.

The top two ovals in Figure 3-2 indicate the original domain of the HRS that for many decades focused on living conditions and economic factors. More recently, the HRS has begun to include biological measurements in an effort to increase its data breadth in this space. These new efforts include chip-based measurements of genetic information for many participants. Of course, the HRS is not unique. Many other studies have been conducted throughout this space. Figure 3-3 provides an overview of sixteen prominent studies that have been hugely impactful. What we believe is critical to note is that no single study has ever attempted to tile this entire space, let alone tile it in single individuals – at a within-subjects level. We thus aim the KHP specifically at that goal: Tiling the space of human behavior in a single population.

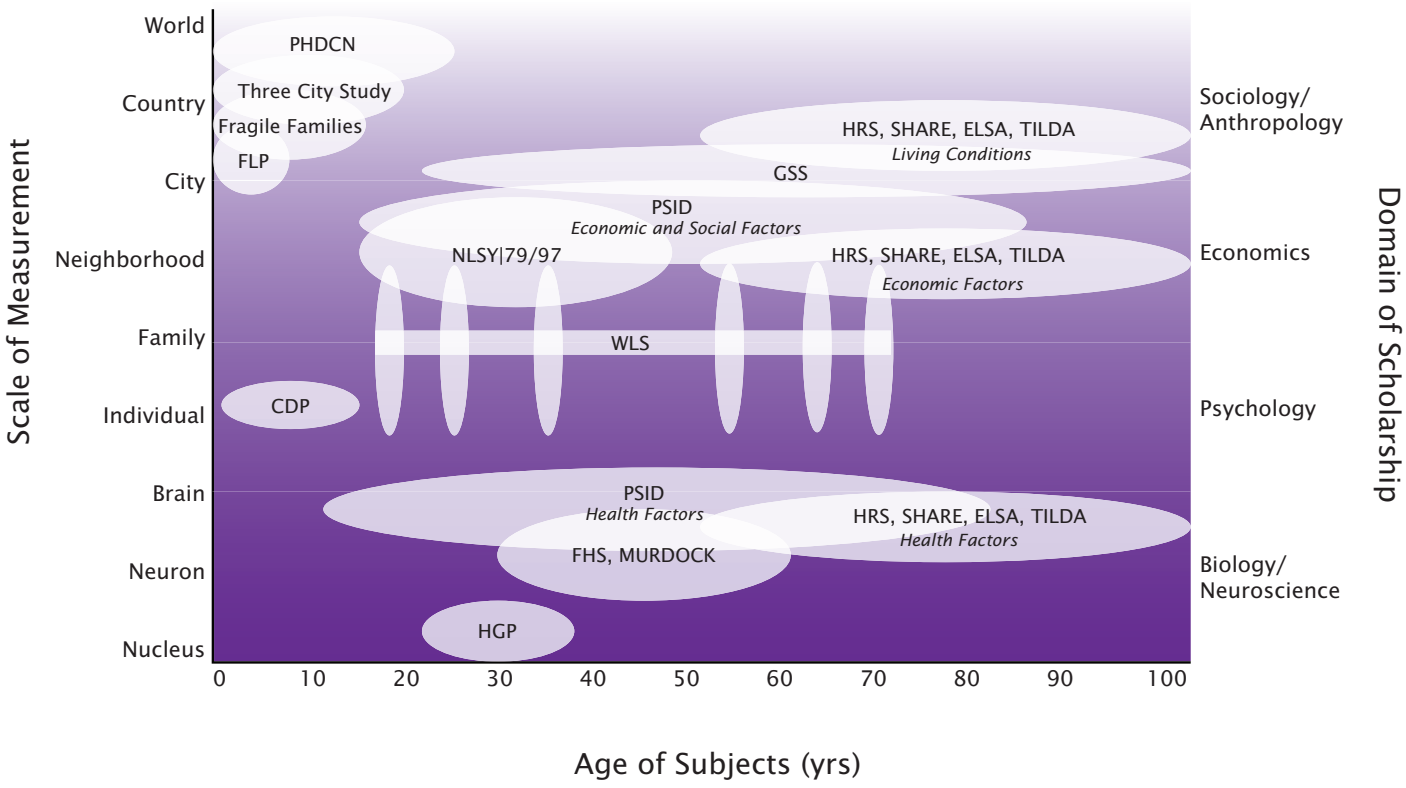


Figure 3-3: *Several Prominent High-Impact Studies.* PHDCN: Project on Human Development in Chicago Neighborhoods, FLP: Family Life Project, HRS: Health and Retirement Study, SHARE: Survey of Health Aging and Retirement in Europe, ELSA: English Longitudinal Study of Aging, TILDA: The Irish Longitudinal Study on Ageing, GSS: General Social Survey, PSID: Panel Study of Income Dynamics, NYLSY: National Longitudinal Survey of Youth, WLS: Wisconsin Longitudinal Study, CDP: Child Development Project, FHS: Framingham Heart Study, HGP: Human Genome Project

2.1 Target Density

To that end, the KHP design process evaluates proposed measurements in a multi-step process designed to develop a complete portfolio of measurements along a number of axes. First and foremost, our goal is to develop a portfolio of measurement that tiles the space of human behavior at the highest possible density.

Figure 3-4 presents a schematic defining the target density of the KHP across this space. Critically, the designed measurement set must provide uniform coverage in areas traditionally associated with a number of academic disciplines ranging from Cell Biology to Anthropology.

2.2 Staged, Conditional Deployment

Note that the time-of-life at which measurements are made will be a critical feature of study design. On the right hand column of Figure 3-4 are measurement domains thought to be largely invariant over the lifespan. These are measurements made once, or very rarely, over the course of the study. Other classes of measurement require regular

periodic sampling. For example, at the bottom of the figure TL: Telomere Length will be sampled at regular intervals across the lifespan. (Our current design allows repetitive physical sampling at 2-3 year intervals, when KHP recruitment and sampling vans revisit families in their homes. More details on the vans and the sampling timeframe they impose can be found in the Study Frame section of this document.) In contrast, C: Cognitive Testing is conducted more frequently during childhood and at a much lower rate during adulthood. That rate increases in late adulthood.

Of course, many measurements must be made in a staged manner across the lifespan – the deployment of tests must reflect the life-stages of our participants. When evidence that an elder participant is becoming senile accumulates, our data collection regimen must change to reflect that change in a participant. As a result, each actual measurement set must be tailored not just the steadily advancing age of each participant but also to changes in the condition and life-status of each participant. Figure 3-5 presents a lifespan trajectory for an example class of measurement. The trajectory curve shows when measurement instruments change and when life events or shocks (such as a change in health status) lead to bifurcations in the instruments required by different subjects. The vertical bars plot ideal measurement times, with rates that often change across the lifespan. The goal of this class of figure is to organize the deployment of measurement instruments in an easily digested fashion. This phased structure assessment for each test must a key feature of the test inventory in the final study design.

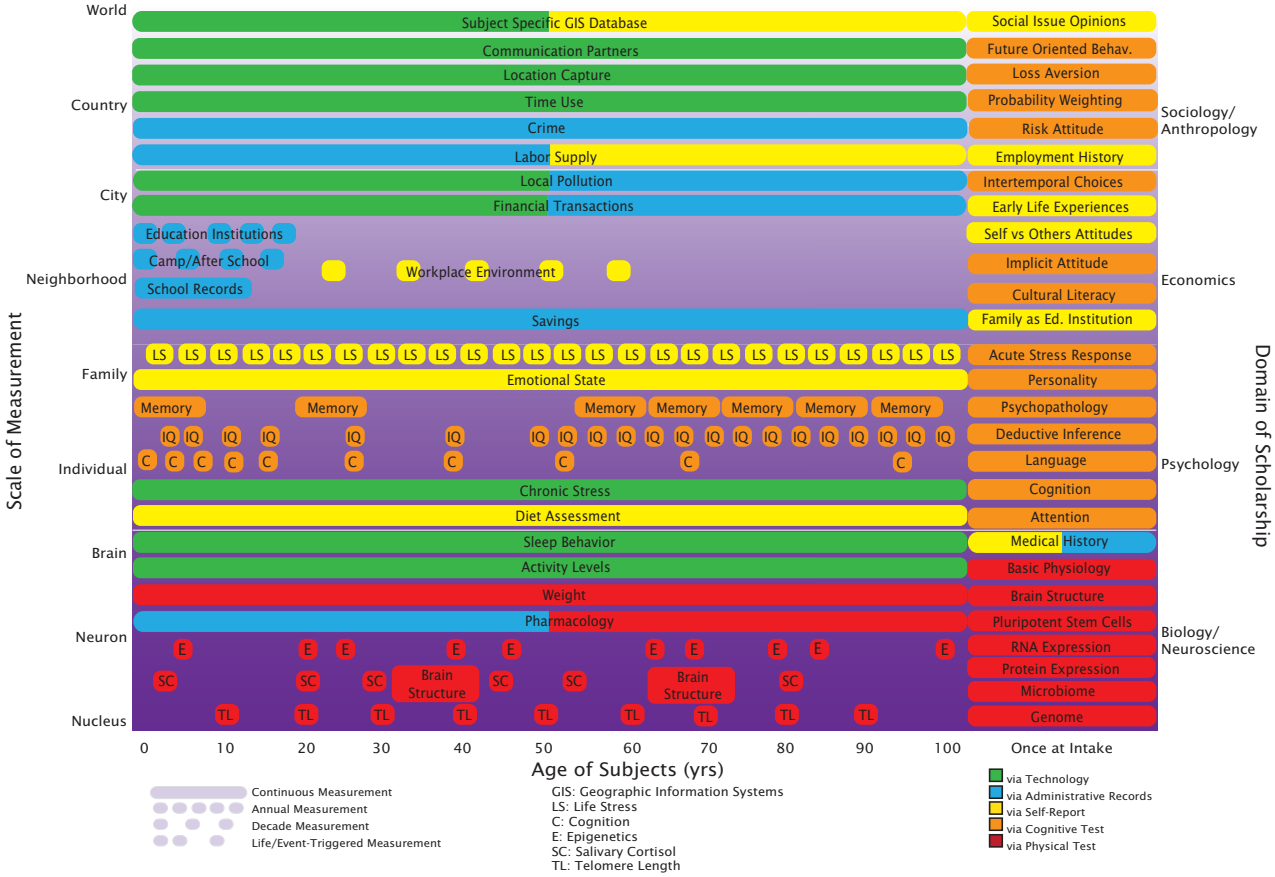


Figure 3-4: *The Target Portfolio of the KHP*

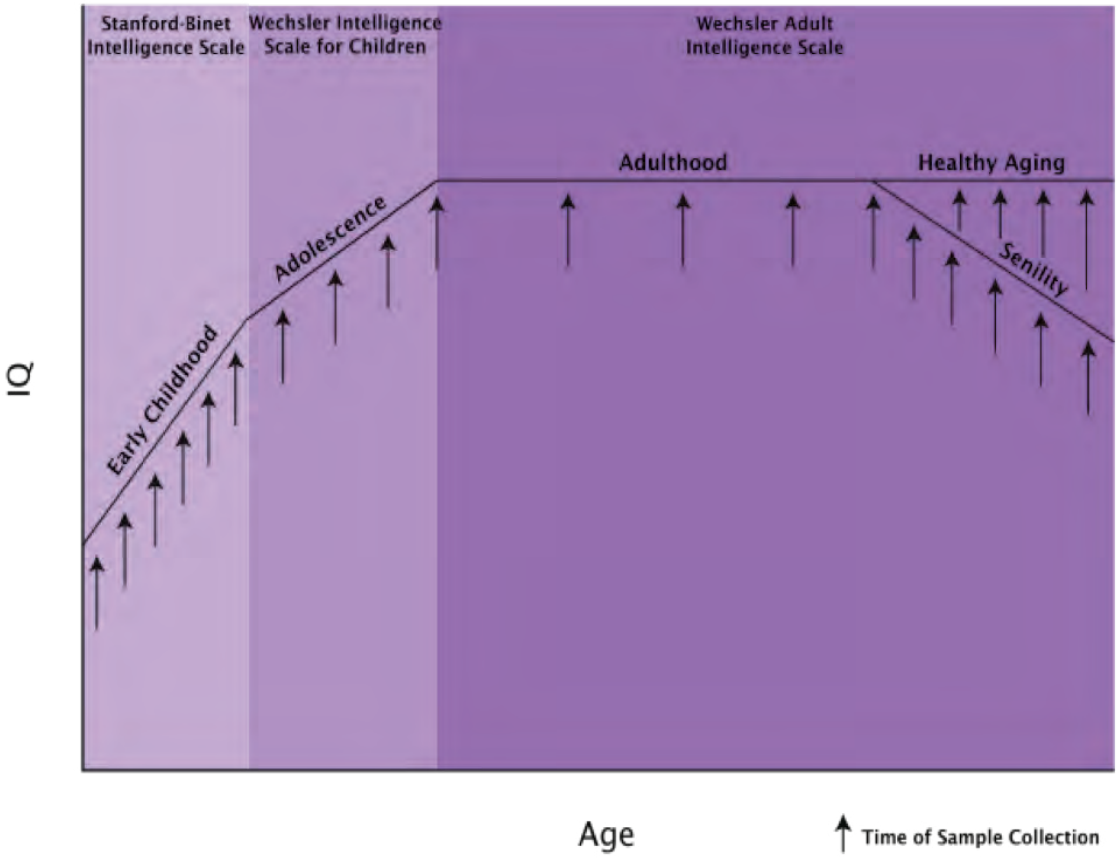


Figure 3-5: *Staged Measurement Deployment*

2.3 Portfolio Must Span Novelty and Cost

A second feature of the measurement portfolio is that it must effectively span the novelty-cost space. Figure 3-6 shows a small set of measures plotted in this space. At the bottom-left are measurements that are well validated and inexpensive. At the top-left are more expensive but highly validated instruments. These form the bedrock of the study at its initial deployment. But because the study is scheduled to run for 20 years, we must expect new measurements to become available that are critical. It is important that these new measures be normed into the KHP database as they become available. To make that possible, it is important that at regular 3-year intervals the current portfolio of tests be assessed in this space. Some tests on the far right of this graph will be dropped in response to this evaluation. Others will have migrated to the left and downwards as they become cheaper and more established. In order to assure that after 20-years our portfolio does not lie exclusively at the lower right corner of this space, new measurements will have to be folded in to the right hand side of this space at regular intervals.

Of course, assessing the importance of utility of a task is critical in determining whether or not it will be included in the KHP portfolio. That determination, guided by the metrics described above, is made by the Measurement and Technology Advisory Council (MTAC) of the KHP. Like the other advisory councils, the MTAC is composed of leading academics who themselves tile the many fields ranging from Cell Biology to Anthropology that the KHP engages. The charge of that board is

first to identify candidate measurements, to assess the importance and desirability of each measurement and then to place that measurement into our analytic framework. Finally, the MTAC must determine whether the KHP has sufficient statistical power to make that measurement useful or valid. This is a critical point for the MTAC to engage for each measurement. Classes of measurements for which the KHP cohort is simply too small to yield meaningful data should be approached with great caution, especially if those measurements are expensive to make. The KHP staff uses that information to propose portfolio structures to the MTAC for evaluation and approval prior to completion of the final study design.

The critical elements of the final portfolio evaluation thus rest on:

1) Tiling the space of human behavior
2) Tiling the age-distribution of human behavior
3) Assessing the cost of a proposed measurement
4) Assessing the novelty-established validity of a proposed measurement
5) Assessing the desirability, usefulness or criticality of a proposed measurement
6) Determining the required sample size to achieve meaningful statistical power with the proposed measurement

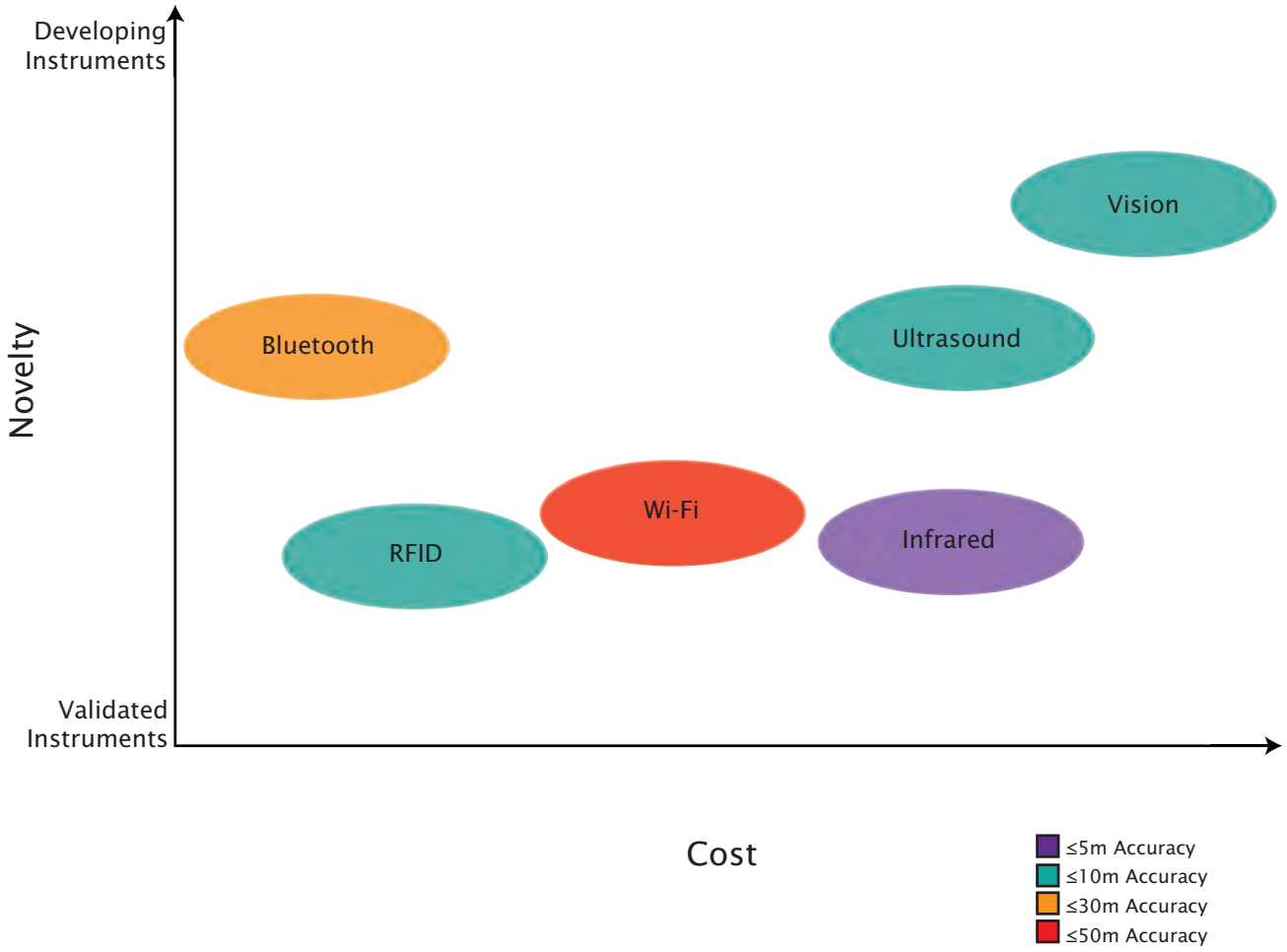


Figure 3-6: *The Novelty Cost Space*

3. Nesting Three Cohorts

In developing the KHP measurement portfolio in accordance with these design approaches, two conflicting goals are immediately highlighted. First, although the largest study of its kind, the KHP cohort is simply not enormous. At roughly 10,000 individuals it does not provide adequate statistical power for answering some kinds of questions. A larger cohort would be better, but of course a larger cohort would be much more costly. Increasing the size of the cohort by a factor of 10, or a factor of 100, is simply impractical at a financial level.

On the other hand, at 10,000 the cohort is so large that many highly desirable measurements located in the top right corner of the cost novelty space are simply infeasible. Structural brain imaging is an example of this kind of measurement. While highly important, structural brain imaging is simply too expensive to undertake with a cohort of 10,000 at the proposed budget of the KHP.

To engage these two issues: that the cohort is too small for some kinds of inquiries and too large for some expensive classes of measurement, we propose embedding the KHP cohort into a three-tiered super-cohort, as shown in Figure 3-7.

Shown in green is the core cohort of the KHP. We propose, however, to identify 250 individuals from the cohort for a set of more expensive testing (yellow circle at the bottom). Testing for this subcohort will include high-value, high-cost measurements that would be impractical at scale.

Shown in blue is the super-cohort. Our vision is that the basic measurement tools that are part of the KHP API can be deployed in the larger New York City community (and perhaps beyond). Interested individuals could become part of the KHP community by simply loading the KHP app onto their smartphones, completing the same demographic questionnaires used in the core sample, and then by participating in all of the app-based data collection undertaken by the core sample.

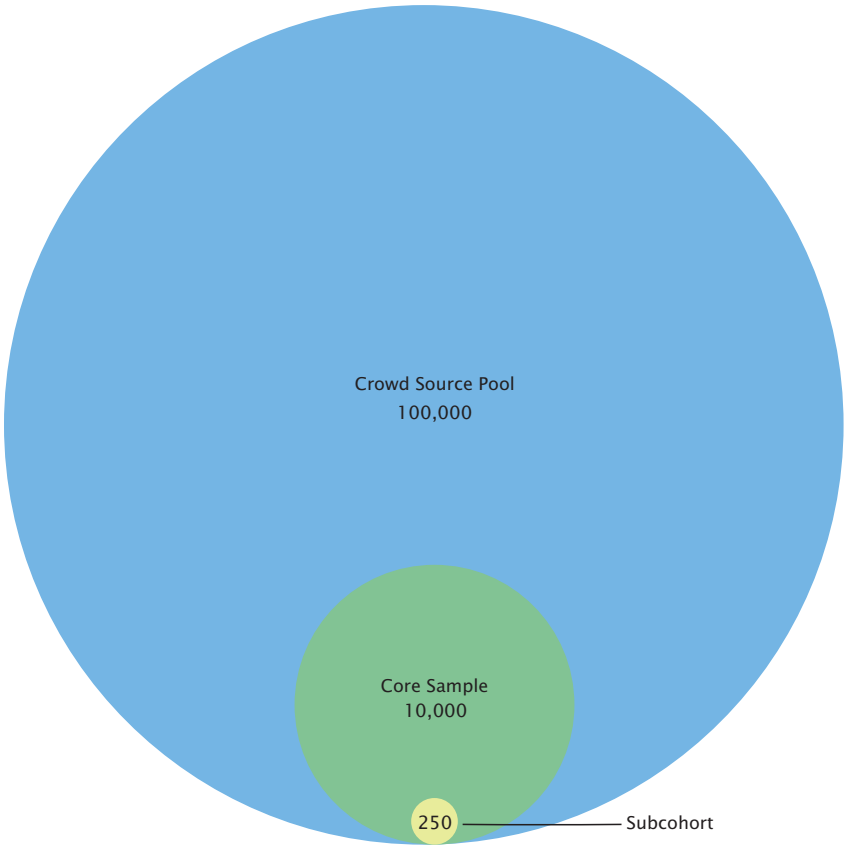


Figure 3-7: *The Nested Cohorts*

This crowd-sourced pool can serve two purposes for the KHP. First, it allows hundreds of thousands of New Yorkers to participate in the KHP at a very intimate level. Second, it provides an enormous amount of data at very low cost – although data that lacks the random cross-sectional structure of the core sample. For many technical or policy questions the crowd-sourced pool is inappropriate, but the crowd-sourced pool will allow citizens of all kinds to contribute to the KHP mission. More details on the crowd-sourced pool, the mechanism for storing data from the pool, for anonymizing that data before it enters the KHP database, and for protecting the security and privacy of crowd-source participants can be found in the Study Frame section of this document.

4. Data Collection and Ingestion

The KHP will need to work with both unstructured and structured data from analog and digital sources. Unstructured analog data, for example, could take the form of hand-written notes by health care professionals. This data set will be scanned and stored for eventual use by automated handwriting analysis systems, except in a limited number of cases. Structured analog data will take the form of standardized test results (for example medical test results or questionnaires), which can be converted into structured digital data with ease using existing technologies. Digital structured data will include input from digital devices, such as smartphones, wristbands and Bluetooth beacons, as well as data sourced from external databases, such as Electronic Medical Reports and NYC GIS. Figures 3-8 and 3-9 illustrate our proposed approach to data ingestion.

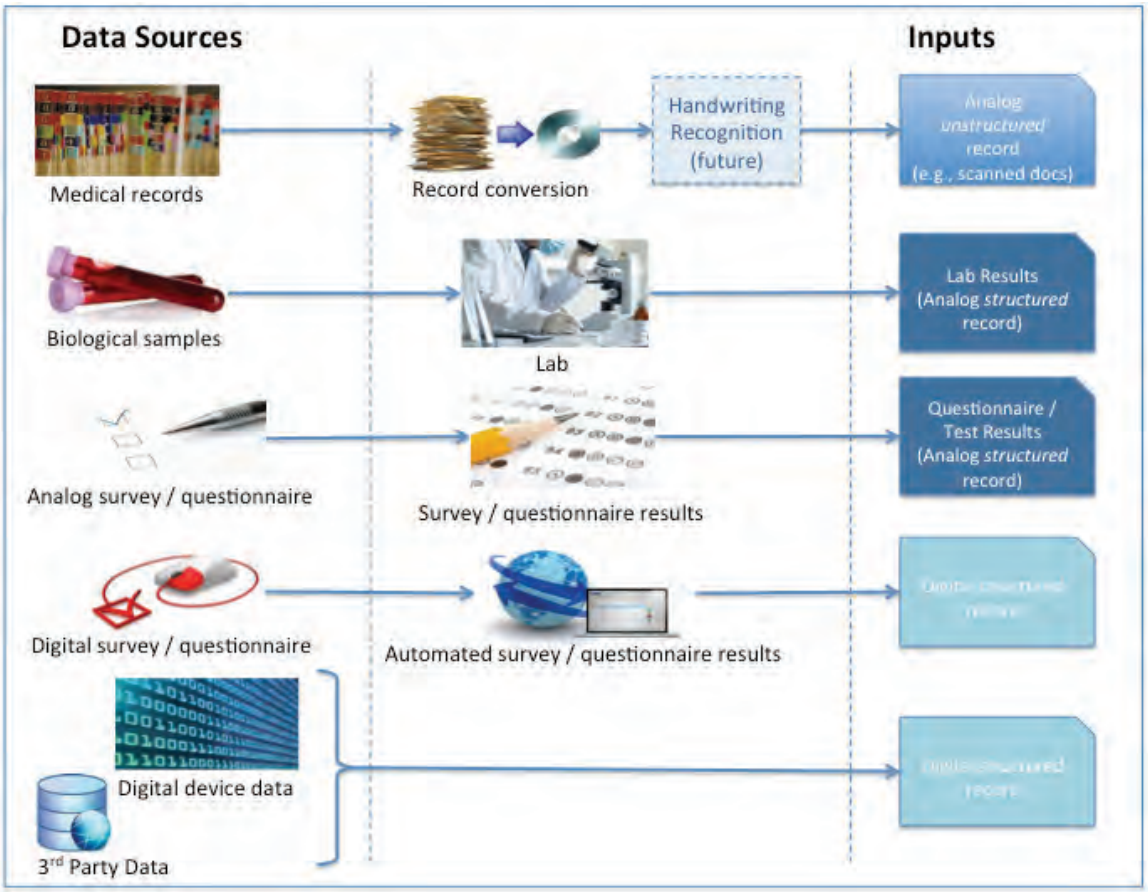


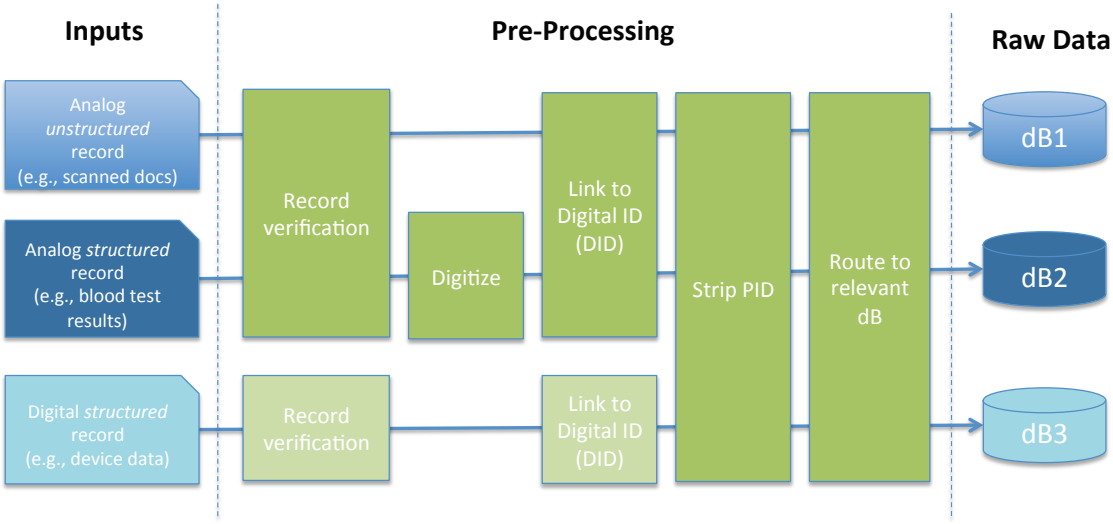
Figure 3-8: *Data Sources to Inputs*

4.1 The KHP App

Smartphones are a key piece of infrastructure for collecting data about study participants. KHP staff will install an application on the phone of each individual above the age of 10 years old (the study will furnish phones and contracts to those who enter the study without a personal smartphone) that has full set of measurement capabilities. The app will have the capability to measure distance to Bluetooth beacons (for child-parent or other social interactions), information about other local devices (e.g., smart home appliances), location data, incoming and outgoing SMS and MMS messages, phone calls and email counts, social media profiling (time spent on these sites, number of updates and number of contacts), a portal for managing the context delivery of questionnaires and surveys and an interface for participants to respond to queries.

4.2 The Data Processing Platform

The project's choices for data processing platforms will be forward-looking in order to adapt to medical and technological advances during the lifetime of the study, which will undoubtedly introduce more accurate measurements of existing data over time, for instance through the use of more accurate air quality sensors. It is also reasonable to expect that new data types would also enter the study, such as blood glucose level measurements from temporary tattoos. At times, new data points can also become available retroactively, for instance when a new blood test may be applied to stored blood samples that had been collected years earlier. In addition, as existing external databases (e.g. NYC GIS) expand their capabilities in the future, the study will have the opportunity to incorporate or link to increasingly more granular third party data. While



Notes:
1. Analog unstructured record (e.g., doctor's notes) assumes that the information will be stored "as is" for future reference or capabilities (e.g., reliable handwriting recognition).
2. Analog structured record assumes reliable automated means of digitizing data, such as results of blood tests, urine tests, etc.
3. Analog records may require some manual oversight / processing in "Record verification" and "Strip PID" steps in the process.
4. Digital unstructured data is not expected at the moment.

Figure 3-9: *The Data Ingestion Process*

the specifics of future data types and impacts to study population cannot be known at the outset of the project, we can make two projections with high confidence: (a) as the project progresses there will be a broader variety of data points, both from study participants and from external sources, very possibly collected at higher frequency and (b) as a result, data volume and processing needs of the project will also continue to grow.

Consequently, KHP study's data architecture will plan for these changes in its design and selection of database platform. Data architecture will incorporate flexibility to define new data points in the future without requiring a major overhaul of existing databases, while maintaining performance of the system overall. Similarly, metadata about collection methodologies are to be incorporated into the data architecture to reflect potential impact of future technologies on data accuracy and sampling frequency. For instance, metadata will include makes & models of sensors, their measurement sensitivities, and the dates when those sensors were in use for each participant. Lastly, a flexible document database under "NoSQL" umbrella is considered for the study.

In order to support the expected growth in processing demands, the Project plans to utilize scalable clustered solutions, such as Hadoop / MapReduce, and evolution of those technologies. All implementation decisions will also need to take into account the costs & benefits of creating a stand-alone solution versus utilizing cloud-based solutions, such as Amazon Elastic MapReduce (EMR) for a Hadoop platform or Amazon RedShift for data warehousing, either end-to-end or for specific functional needs. In these decisions the project will seek state of the art solutions that offer proven security, reliability and performance.

4.3 Record Validation

Record validation is an area of particular focus in the KHP study to ensure that incoming data will be complete and accurate. Records in each data stream are to be validated in two stages, taking into account all available information both from the particular data stream and other data streams.

The first stage of record validation is planned to be point-wise validation, which will ensure that observed data for a particular data point (variable) arrives at the expected frequency and volume, in expected formats, and it is within the valid range of values for the variable. For instance, geo-location data from participants' smartphones should include valid latitude and longitude values, and it should arrive every few minutes, taking into account cases when the smartphone may be without cellular or WiFi coverage, or without power.

The second stage of record validation is planned to ensure that each data point will be accurate through three mechanisms: cross-validation, predictive checks and error correction. The first mechanism, cross-validation, will leverage other data streams, when possible, to crosscheck individual records. For instance, when a participant is at home, presence data from Bluetooth based sensors should be consistent with geo-location data from the participant's smartphone. The second mechanism, predictive checks, will use historical data for the particular data stream and from other relevant data streams to predict expected value for a data point, and compare the observed value to the predicted value, in order to highlight potential "exceptions." For instance, a short-term prediction for the expected geo-location of a participant could leverage the most recent geo-location and velocity data to predict a specific participant's location within the next few minutes, in order to highlight potential issues with reported geo-location data. A longer-term prediction that uses longer-term behavioral history of the participant could identify a shift from historical trends to explore whether an unexpected set of values is due a potential error or change in behavior. For example, a change in morning commute routine one day could indicate the start of a change in employment, a transient change in behavior (e.g. due to suspension of subway service), or a technical problem with geo-location module in the smartphone. The third mechanism, error correction, will utilize conservative error-correction measures to compensate for records with potential

errors and for missing records. This mechanism will only update an erroneous or missing record when it has high confidence; otherwise, it will remove potential errors from input and leave missing records untouched. Error correction will be applied to all possible data streams. For example, it can be applied to a blood test in the following manner: when lipid values in a blood test do not align with expected range of values, given the physiology and medical history of the participant, one possible error correction approach would be to order a second test to eliminate a potential error.

4.4 Data Quality and Volume

Close consideration is given to data quality within each information domain of the project. For instance, to collect data about participants' movements and activities at a reasonable cost, the KHP plans to leverage mature and widely used data collection platforms, such as smartphones (iOS and Android) and activity trackers, while developing custom "apps" designed around the needs of the study. In collecting bio-medical and psychological data, the study group will work with highly qualified agencies and scholars to ensure repeatable and reliable test results. Lastly, the KHP will establish proactive monitoring of data in-flow in order to identify and resolve potential issues before they can impact the overall quality of data. Such issues may be due to factors such as technical problems, temporary outages (e.g. battery drainage) or participants' failure to follow study requirements.

The KHP plans to capitalize on the current technologies that enable the rapid acquisition of substantial amounts of electronic data. Automated data collection enhances the comprehensiveness of the available data, but it can also help with verification of data as it provides multiple views of any particular event. For example, location data for an individual should coincide with purchase information at a store at that same location. Of course, there will be particular areas where automated data collection will be difficult, if not impossible, and we will have to rely on proxies or more granular level data when these problems arise.

While it is unlikely that people will be willing to keep detailed food diaries over long stretches of time, diaries for shortened durations combined with detailed purchase information from restaurants and grocery store receipts can still provide meaningful data about diet that transcend these kinds of problems. Cash purchases are another area in which indirect measures are required to make inferences. However, in all cases, the KHP study plans to implement a detailed and multifaceted data collection effort with current electronic technologies in order to estimate proxies for economic and behavioral measurements that have not previously been studied quantitatively at scale by any group.

One of the KHP's key objectives is to identify early predictors of health outcomes, even before such outcomes can be diagnosed, such as identifying subtle changes in behavior that may indicate eventual onset of a disease (e.g. Alzheimer's or Parkinson's). A key determinant of data quality for a time-series analysis of this kind is sampling frequency, where high sampling frequency enables researchers to observe small or transient changes that may ultimately turn out to be reliable predictors. Consequently, the KHP study plans to collect data at high frequencies, depending on the expected rate of change in each information domain, to obtain a "high definition" view of study participants' lives.

The KHP also recognizes the importance of gathering detailed historical information on medical and environmental experiences. The study plans to collect the following information from participants: complete genetics, complete microbiomes, standard physical examinations (including lab work and psychological examinations), social networks/safety nets, geo-location data, activity tracking (by band or by smartphone), sleep tracking, hair-analysis, urine-analysis and pharmacy records for assessing drug use, direct measures of stress levels, environmental quality (air, noise, chemical exposure via the standard silicone wristband approach), detailed purchasing data (particularly food types), work experience (most Americans spend as much time at work as at home, and the effects of heavy physical demands, a sedentary setting or chronic stress are

likely to have substantial influences on health outcomes), detailed structured and unstructured medical records (including ongoing exams and treatments) and social services records (disability, Medicaid, Medicare, SNAP).

4.5 The Data Collection Process for Participants

An Overview of the Data Collection Process:
Pre-enrollment (0.5-1 hr)

- Full consent process

Enrollment (3-4 hrs)

- Physical sample collection
- Medical history
- Psychological testing that requires a PhD psychologist (SCID, IQ)
- Tech installation (including enrollment in any online services)

Forward going weekly/monthly/bi-annual pings to mobile phones to complete surveys

- Time budgeted to avoid overwhelming subjects
- Gamified – appropriately targeted to kids and to adults (e.g. Social network construction instrument)
- Potentially to include small scale experiments for additional data collection (approved by board of directors, but not paid for by HUMAN project). Potential examples include activity trackers or exposome silicone bracelets.

Annual or biennial re-capture

- Physical exam

4.6 The Time Budget

At all stages of the data collection process, time is of the essence – the desire to collect as much data as possible must be weighed against the willingness of participants to spend time contributing to the study. Table 3-1 provides notional time allocations for the various intake stages.

We will organize data collection so that any processes requiring physical samples or the presence of trained personnel are done during the enrollment process, which we expect could last about to three to four hours. Biological sampling is also limited by how much material (blood, urine, saliva and stool) an individual can provide during a single contact. Technology installation can run in parallel with data collection, with the exception of instructions necessary for future operations. Details of the enrollment process can be found in the Study Frame Design chapter, and the full list of psychological exams and tests run on biological samples can be found at the end of this document.

Although we expect that the bulk of the data will be collected automatically following the enrollment process, there will still be some forms of data collection for which the subjects will have to be active participants (e.g. social network self-reports or surveys about major life events). We propose to develop a time budget for participants on a weekly, monthly and annual basis that balances our data collection needs with the amount of time that participants can reasonably be expected to give – and be compensated for. This budget will include a range of time demands from shorter, more frequently gathered items like surveys to longer and more invasive procedures like updating biological samples. Frequent events might be gathered by via mobile phones on a weekly basis, more detailed procedures as rarely as annually or biennially.

All forms of data collection requiring active participation will need to be game-ified in order to encourage compliance. The KHP mobile application will be designed to present these "games" (questionnaires and surveys) at a time of day that is convenient for each individual. Table 3-2 lists the different technologies that will be installed and/or provided to study participants for data collection.

Many kinds of data collection may be impossible across the entire study population, either because the data collection technology may not be stable across years or because of the high cost of data collection. When that is true, it may be advisable to design methods for intermittent data collection. It

might, for example, be interesting to deploy advanced physiological monitors or other tools for 3-week periods to randomly selected study participants. One example of this would be to deploy environmental toxin monitors for 3-week periods randomly through the study population such that 50 subjects are monitored in this regard at all times.

In addition to these regularly scheduled data collections, we also envision performing life event-triggered sampling. In the most straightforward case, we can follow up with participants any time that they report an event on one of the regular traumatic life event surveys. However, we could also trigger data collection protocols on large changes in the incoming data streams such as radical changes in the nature of a subject's social network or physical location. The development of sample triggers is proposed as a major focus of the CD2 process.

The overall study would also benefit if independent investigators with their own funding were offered the opportunity to use new data gathering technologies within the study frame. A cardio-vascular study group, for example, might be offered the opportunity to recruit subjects for wearable EKGs for a limited period. The study management team will have to include a mechanism for reviewing proposals of this kind.

Table 3-1: *Time Budget for Subjects During Intake*

Activity	**Duration**	**Performed by**
Intake Interview	60 minutes	Enrollment specialist
Physical examination & biological sample	45 minutes	Nurse
Mini International Neuropsychiatric Interview	30 minutes	Psychologist
Wechsler Adult Intelligence Scale (WAIS) or WISC for Children	30 minutes	Psychologist
Technological set-up for subject (KHP app, devices)	30 minutes	IT
Technology set-up in the home	30 minutes	IT
Total	3 hours 45 minutes	

Table 3-2: *Technology to be Deployed to Subjects*

Technology Item
KHP App on all iOS devices (iPhone, iPad)
KHP App on all Android devices (smartphones, tablets, phablets)
Bluetooth beacons in each room of the apartment, in fixed locations
Bluetooth-enabled clothing for children in the study
Activity tracker for adults
Base station in the home (to serve as in-home collection point for Bluetooth data, and to upload it to KHP servers)
Silicone wristband (to monitor exposure to chemical compounds)

4.7 Focus Groups

Objectively, the proposed measurements span the spectrum of physical and personal invasiveness. However, opinions about what constitutes an invasion of personal privacy are quite diverse and idiosyncratic – some are willing to reveal detailed drug taking behavior, but recoil from discussing political party registration (even though this is a matter of public record), others feel that web search history should be more protected than location information, and in either case, there are people with opposing views. It will be important to broadly sample opinions from our potential subjects about what they perceive as being invasive, and why, so that we can formulate strategies for addressing these concerns with our Public Outreach and Privacy Councils. We propose to implement focus groups in the coming months to begin this process.

5. Overview of Datatypes

Recall that the KHP database rests on three classes of information: 1) Data about the environment in which our participants live and work, 2) Physical and Biological Data drawn directly from our participants and 3) Digital data gathered at a subject-specific level from a number of interrelated sources. To achieve this level of density our datasets will include not only data that we gather directly, but data from other sources that has been integrated into the KHP database.

A number of active projects by New York City offer the opportunity for added power to the KHP study in this way by cross-referencing data from study participants with data from these projects. One of these is a recent project by the NYC Health and Hospitals Corporation (HHC) to modernize the electronic health records system in cooperation with 5 major medical systems in the NYC area, which includes NYU. It is expected that by 2017, electronic medical records from across all NYC HHC patient care facilities, including hospitals, long-term care facilities, diagnostic treatment centers and community based clinics will be fully integrated. The 5 major medical centers should achieve full integration into this system slightly later. The result will be an electronic medical record system of unparalleled quality in the United States. The NYC HHC, alone is the largest municipal health system in the country, and treats about 1.4 million patients a year, including a large proportion of the uninsured. The HHC medical record system conjoined with a consortium of New York's other large-scale medical providers with should yield the most comprehensive electronic medical record system in the United States. Access to that database provided

by conducting the study inside New York City will make the proposed dataset of exceptional use to medical researchers.

Another data resource that is available to a New York-based study is the Statewide Planning and Research Cooperative System (SPARCS). The SPARCS database contains individual-level detail on patient characteristics, diagnoses and treatments, services and charges for each hospital inpatient stay and outpatient visit (it includes ICD-9 codes and data on ambulatory surgery, emergency department, and outpatient services); and each ambulatory surgery and outpatient services visit to a hospital extension clinic and diagnostic and treatment center licensed to provide ambulatory surgery services. In addition, New York City maintains a relatively new prescription-drug monitoring program registry for all controlled substances, which contains individual level records.

Another significant external data source available only in New York at this time (although Chicago is rapidly also developing such a system) will be the Geographic Information System (GIS) that is currently being built by New York University as a resource for the study of New York by CUSP. It is important to remember that this database will provide a multi-layered view of New York City, capturing information about air pollution, electricity use, garbage volume, emergency service requests, noise complaints and even parking tickets issued and the individuals to whom they are issued. Overlaying this information with data that is collected from study participants will enable investigations of interactions between the environment, behavior and biology to understand how these factors turn a predisposition for something like heart disease, diabetes or depression into pathology in only a subset of those who have the vulnerability.

Table 3-3 below presents a more detailed overview of the types of information and data sources that will be part of the study. A comprehensive list of each individual measurement is included as a detailed table that forms the final segment of this chapter.

Table 3-3: *KHP Measurements*

Information Domain	Inputs into KHP Study	Data Sources
Demographics	Demographic information about participant household and individual members of the household, such as age, gender, ethnicity.	• Participant questionnaire • Supporting documentation, e.g., birth certificate, driver's license or passport.
Home Environment	Information about housing space, presence of toxins, air quality, ambient noise level, water and energy use.	• Participant questionnaire • Building information • Survey and measurements by KHP field team • Sensors for air quality, ambient noise • Utility records

Neighborhood Baseline	Census data, education quality and the crime statistics for the neighborhood in which the participant lives, such as demographic composition, median income, school ratings, emergency service requests and crime statistics.	• NYC public data sets on census, education, law enforcement, public service and GIS • NYU CUSP databases
Bio-medical	Information about each participant's medical and dental history, physiology, biochemistry, whole genome, complete microbiomes and complete pharmacological use profiles.	• Physical exam (weight, height, BMI, resting heart rate, blood pressure) • Blood sample (for genetics and blood chemistry) • Urine sample (for toxicology) • Saliva sample (for oral microbiome, genetics and stress measurement) • Hair sample (for toxicology and chemical exposure) • Stool sample (for gut microbiome) • Electronic Medical Records (EMRs), doctor's notes, dentist records, hospitalization history • Health insurance records • NYS database on health data (SPARCS) • NYS database on prescriptions • In limited numbers: Silicone wristbands (for chemical exposure) • In limited numbers: functional MRI, electroencephalogram (EEG) and electrocardiogram (EKG) for more invasive study of core set of participants.
Diet and Health	Information about each participant's diet, use of alcohol, tobacco and other substances.	• Participant food diaries (for limited durations, repeated at intervals) • Financial transaction records, mined for food & health related purchases as well as for wealth and prosperity measurements
Psychological	Information about participants' mental health, persistent strain, personality attributes, level of cognitive function and risk preferences.	• Structured interviews of participants by trained professionals to include assessments of mental health and intelligence • Self-administered tests on smartphones and tablets

Educational	Information about participants' formal and informal educational history (e.g. number of books in the home) and progress of current education.	• Participants' educational records and extra-curricular activity records • Survey of participants' homes by KHP field team • NYC Department of Education databases on school rankings and progress of individual students • Purchased data on college-level performance
Occupational	Information about participants' occupational history and progress of their occupation / career during the study time frame.	• Participants' curriculum vitae (oral or written) • Participants' W-2 records, unemployment insurance applications
Activity	Information about the times and duration of different activities, such as sleep, commute / travel, work / school, exercise, entertainment, socializing and screen time, as measured by wearable technologies, smartphone apps and presence detection systems.	• Smartphone app (for location, activity and socializing data) • Wearable activity trackers • Bluetooth-based presence sensors in participants' home • Smartphone / tablet app for social media and digital contacts • NYC GIS database
Family Interactions	Information about the frequency and duration of interaction between parents and children in the home. Information about the level of care giving to family members by family members.	• Participant questionnaire • Bluetooth-based presence sensors in participants' homes.
Financial	Information about participants' sources of income, major assets & liabilities, categories of expenses, savings and retirement planning activities. Detailed purchase data to the level of all individual purchases, grocery purchases at the level of individual items, prescription drug co-pay data, alcohol purchases, tobacco purchases, etc.	• Participant questionnaire • W-2's • Title and ownership documents for key assets • Bank records • Credit card and debit card records • Loan records • Public Assistance records (e.g., SNAP) • Retirement planning account information (e.g., pension, 401k) • Rental agreements, mortgage records

Interactions with Law Enforcement	Information about participants' interaction with law enforcement agencies (either as victims or potential culprits), and with the criminal justice system.	• Participants' call history • NYC Police Department databases on 911 & 311 calls, stop & frisk activity, arrest records • NYC District Attorney's records on case histories

6. Proposed Instruments and Measurements

6.1 Demographics

6.1.1 Household Demographics

The following information will be collected about the Household Composition:

Table 3-4: *Demographic Measurements*

No	Data Element	Frequency
1	Number of people living in the home	Annually
2	Number of generations living in the home	Annually
3	Number of adults (19+)	Annually
4	Number of teenagers (ages 12 – 18)	Annually
5	Number of children (<12)	Annually
6	Number of senior citizens in the home age (age 60+)	Annually
7	Primary language spoken at home	Annually
8	Secondary language(s) spoken at home	Annually

6.1.2 Individual Participant Demographics

The following information shall be collected from each household member. Information about each member of the Household will be held at lower granularity to preserve anonymity.

Table 3-4: *Demographic Measurements (continued)*

No	Data Element	Frequency
1	Year of Birth	Once
2	Gender	Once
3	Sexual Identity	Once
4	Ancestry	Once
5	Ethnicity	Once
6	Marital Status	Once

7	Head of Household (Y / N)	Once
8	Place of Birth	Once
9	Nationality	Once
10	Year of naturalization (If nationalized citizen)	Once
11	Native language	Once
12	Other languages spoken	Once
13	Relationship to each member of the Household	Once

6.2 Home Environment

Information about participant's dwelling shall be collected by the KHP Field team. Information about air quality and ambient noise levels will be collected via sensors in the home. Dwelling information and sensors will be updated when the family moves to a new location.

Table 3-5: *Home Environment Measurements*

No	**Data Element**	**Frequency**
1	Total interior square footage	Intake & each move
2	Number of bedrooms	Intake & each move
3	Number of full & half bathrooms	Intake & each move
4	Year when the building was built, is the information exact / estimate	Intake & each move
5	Primary heating system & technology (e.g., central heating with steam, per-room electric heating) and locations	Intake & each move
6	Air-conditioning system(s) & technology (central A/C, per-room wall-unit) and locations	Intake & each move
7	Means of food preparation (e.g., stove top, oven, microwave, hotplate); capacity (e.g., number of burners, cu-in for microwave)	Intake & each move
8	Approximate size of refrigerator	Intake & each move

9	Make, model & energy consumption of appliances: Fridge Microwave Electric stovetop & oven Electric heaters TVs Other	Intake & each move

6.2.1 Chemical Exposure

Chemical exposure will be measured using silicone wristbands, which will be worn by approximately 300 participants every month. At the end of the month, study participants will mail the silicone wristbands to a KHP lab, which will perform a spectrogram to determine the levels of chemicals the wristband—and thus participant—are exposed to.

Each month a different group of 300 participants will be asked to wear the silicone wristbands. However, any participant in previous groups who wishes to continue to participate in this part will also be provided with a new silicone wristband.

Table 3-6: *Chemical Exposure Measurements*

No	Data Element	Frequency
1	Spectroscopy of silicone band	Every 3 years

6.3 Medical and Dental Status

6.3.1 Physiology

Table 3-7: *Physiology Measurements*

No	Data Element	Frequency
1	Height	Every 3 years
2	Weight	Every 3 years
3	Body Mass Index (BMI) - calculated	Every 3 years
4	Standing waist circumference	Every 3 years
5	Arm circumference	Every 3 years
6	Resting heart rate	Every 3 years
7	Resting systolic blood pressure	Every 3 years

8	Resting diastolic blood pressure	Every 3 years
9	Temperature	Every 3 years
10	Pulse oxygen	Every 3 years
11	Exhaled gas	Every 3 years

6.3.2 Medical Status

Medical status will be obtained through two different means:

a. via Electronic Medical Records (EMRs) whenever possible. Otherwise via reaching out to subject's physician, with subject's consent.
b. via Statewide Planning and Research Cooperative System (SPARCS) database, which contains individual-level detail on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient visit (it includes ICD-9 codes and data on ambulatory surgery, emergency department, and outpatient services); and each ambulatory surgery and outpatient services visit to a hospital extension clinic and diagnostic and treatment center licensed to provide ambulatory surgery services.

All data will be updated on an annual basis.

6.3.3 Pharmacological Status

Pharmacological data history will be obtained through three different means:

a. via Electronic Medical Records (EMRs) whenever possible. Otherwise via reaching out to subject's physician, with subject's consent.
b. Via New York City prescription drug monitoring database, which contains individual level records for all controlled substances.

All data will be updated on an annual basis.

6.3.4 Dental Status

When possible, dental history will be obtained by contacting subjects' dentists, with consent of the subject.

All data will be updated on an annual basis.

6.3.5 Blood Analyses

Several blood analyses will be performed at 3-year intervals.

Table 3-8: *Blood Analyses*

Blood Chemistry Screen (SMA-20)

No	Data Element	Frequency
1	Albumin	Every 3 years
2	Alkaline Phosphatase	Every 3 years
3	Alanine Aminotransferase (ALT)	Every 3 years
4	Aspartate Aminotransferase (AST)	Every 3 years
5	Bilirubin (total and direct)	Every 3 years
6	Blood Glucose	Every 3 years
7	Blood Urea Nitrogen	Every 3 years
8	Calcium (Ca) in Blood	Every 3 years
9	Carbon Dioxide (Bicarbonate)	Every 3 years
10	Chloride (Cl)	Every 3 years
11	Cholesterol and Triglycerides Tests	Every 3 years
12	Creatinine and Creatinine Clearance	Every 3 years
13	Gamma-Glutamyl Transferase (GGT)	Every 3 years
14	Lactate Dehydrogenase	Every 3 years
15	Phosphate in Blood	Every 3 years
16	Potassium (K) in Blood	Every 3 years
17	Sodium (Na) in Blood	Every 3 years
18	Total Serum Protein	Every 3 years
19	Uric Acid in Blood	Every 3 years

Complete Blood Count

No	Data Element	Frequency
1	White Blood Cells (WBC)	Every 3 years
2	Red Blood Cells (RBC)	Every 3 years
3	Hemoglobin / A1C	Every 3 years

4	Hematocrit or packed cell volume (PCV)	Every 3 years
5	Mean corpuscular volume (MCV)	Every 3 years
6	Mean corpuscular hemoglobin (MCH)	Every 3 years
7	Mean corpuscular hemoglobin concentration (MCHC)	Every 3 years
8	Red blood cell distribution width (RDW): the variation in cellular volume of the RBC population.	Every 3 years
Differential / Platelet:		
9	Neutrophil granulocytes	Every 3 years
10	Lymphocytes	Every 3 years
11	Monocytes	Every 3 years
12	Eosinophil granulocytes	Every 3 years
13	Basophil granulocytes	Every 3 years
14	Manual count	Every 3 years
Platelet:		
15	Platelets	Every 3 years
16	Mean platelet volume (MPV)	Every 3 years

Lipid Profile

No	**Data Element**	**Frequency**
1	Total Cholesterol	Every 3 years
2	High density lipoprotein (HPL)	Every 3 years
3	Low density lipoprotein (LPL)	Every 3 years
4	Triglycerides	Every 3 years

Serum Heavy Metals

No	Data Element	Frequency
1	Arsenic	Every 3 years
2	Lead	Every 3 years
3	Mercury	Every 3 years
4	Cadmium	Every 3 years
5	Chronium	Every 3 years

Viral Infections

No	Data Element	Frequency
1	Immune Markers	Every 3 years
2	Hepatitis (anti HBs)	Every 3 years
3	Hepatitis A, B, C, D, E	Every 3 years
4	Herpes 1 & 2 antibody	Every 3 years
5	Human immunodeficiency virus (HIV) antibody	Every 3 years
6	Human papillomavirus (HPV)	Every 3 years
7	Measles / varicella / rubella	Every 3 years

Blood Toxicology Test

No	Data Element	Frequency
1	Alcohol (including ethanol and methanol)	Every 3 years
2	Amphetamines (such as Adderall)	Every 3 years
3	Barbiturates	Every 3 years
4	Benzodiazepines	Every 3 years
5	Methadone	Every 3 years
6	Cocaine	Every 3 years
7	Opiates (including codeine, oxycodone, heroin)	Every 3 years
8	Phencyclidine (PCP)	Every 3 years
9	Tetrahydrocannabinol (THC)	Every 3 years

6.3.6 Urine Analyses

Several urine analyses will be performed at 3-year intervals.

Table 3-9: *Urine Analyses*
Urine Trace Materials

No	Data Element	Frequency
1	Arsenic	Every 3 years
2	Antimony	Every 3 years
3	Barium	Every 3 years
4	Beryllium	Every 3 years
5	Cadmium	Every 3 years
6	Cesium	Every 3 years
7	Cobalt	Every 3 years
8	Lead	Every 3 years
9	Molybdenum	Every 3 years
10	Platinum	Every 3 years
11	Thallium	Every 3 years
12	Tungsten	Every 3 years
13	Uranium	Every 3 years

Urine Toxicology

No	Data Element	Frequency
1	Arsenic	Every 3 years
2	Antimony	Every 3 years
3	Barium	Every 3 years
4	Beryllium	Every 3 years
5	Cadmium	Every 3 years
6	Cesium	Every 3 years
7	Cobalt	Every 3 years

8	Lead	Every 3 years
9	Molybdenum	Every 3 years
10	Platinum	Every 3 years
11	Thallium	Every 3 years
12	Tungsten	Every 3 years
13	Uranium	Every 3 years

Trace Pesticides in Urine

No	Data Element	Frequency
1	Malathion dicarboxylic acid	Every 3 years
2	*para*-Nitrophenol	Every 3 years
3	3, 5, 6-Trichloro-2 pyridinol	Every 3 years
4	2-Isopropyl-4-methyl-6-hydroxypyrimidine	Every 3 years

6.3.7 Hair Tests

Several hair analyses will be performed at 3-year intervals.

Table 3-10: *Hair Tests*
Hair Follicle Substance Test

No	Data Element	Frequency
1	Cocaine (Cocaine & Benzoylecgonine)	Every 3 years
2	Marijuana	Every 3 years
3	Opiates (Codeine, Morphine & 6-Monacteyl Morphine),	Every 3 years
4	Methamphetamine (Methamphetamine/Amphetamine & Ecstasy),	Every 3 years
5	Phencyclidine (PCP).	Every 3 years
6	Cotinine (tobacco use)	Every 3 years

Hair Cortisol Level

No	Data Element	Frequency
1	Cortisol Level	Every 3 years

Hair Trace Materials

No	Data Element	Frequency
1	Arsenic	Every 3 years
2	Antimony	Every 3 years
3	Barium	Every 3 years
4	Beryllium	Every 3 years
5	Cadmium	Every 3 years
6	Cesium	Every 3 years
7	Cobalt	Every 3 years
8	Lead	Every 3 years
9	Molybdenum	Every 3 years
10	Platinum	Every 3 years
11	Thallium	Every 3 years
12	Tungsten	Every 3 years
13	Uranium	Every 3 years

6.3.8 Genome and Epi-Genetics

During intake a part of the subject's blood sample will be used to sequence to full genome.

Epigenetics will be captured every 3 years thereafter.

Table 3-11: *Genome and Epigenetics*

No	Data Element	Frequency
1	Full genome sequence (3M base pairs)	Once (at intake)
2	Telomere length	Every 3 years

3	Immune markers	Every 3 years
4	Methylomics – MBD Sequencing	Every 3 years
5	Histone modifications	Every 3 years
6	Chromatic remodeling	Every 3 years

6.3.9 Microbiome

Microbiome will be measured from subject's saliva and stool every 3 years. Stool sample will be collected using the same methodology as uBiome, with toilet paper and swab.

Table 3-12: *Microbiome*

No	Data Element	Frequency
1	Buccal microbiome	Every 3 years
2	Gut microbiome	Every 3 years

6.3.10 Additional Tests for Subcohort

KHP will gather the following measurements from the subcohort only to obtain more detailed information on those subjects:

Table 3-13: *Subcohort Tests*

No	Data Element	Frequency
1	EKG	Every 3 years
2	fMRI	Every 3 years
3	Magnetoencephalography (MEG)	Every 3 years
4	Diffusion Tensor Imaging (DTI)	Every 3 years
5	Cortical thickness	Every 3 years

6.4 Psychological

In addition to gathering information on subjects' psychiatric status through data from medical and pharmacological records and through hair cortisol levels, subjects will be evaluated for psychological status against DSM-V disorders at intake and at 3-year intervals thereafter. Subjects' IQ will also be measured during intake, and at 6 – to 12- year intervals thereafter.

Once the subjects are in the study, additional personality attributes will be measured using standard questions in a validated gamified format leveraging the KHP app. Such questionnaires include Cognitive Battery of the NIH Toolbox, "Face Perception Test", as well as indicators of personal outlook, perceived social status, life stress, risk profile and personality inventory of each subject.

Table 3-14: *Psychological Measures*

No	Data Element	Time Budget (mins)	Frequency
1	Mini International Neuropsychiatric Interview (MINI) & MINI Kid Interview	30	Every 3 years
2	Wechsler Adult Intelligence Scale and Wechsler Intelligence Scale for Children	30	Every 6 – 12 years
3	Biomarker of stress – cortisol level from hair (noted above)	5	Every 3 years
4	NIH Toolbox – Cognitive Battery	37	TBD
5	Face perception test	7	TBD
6	The MacArthur Scale of Subjective Social Status	3	TBD
7	Stress and Adversity Inventory (STRAIN)	35	TBD
8	Self- Esteem Scale	5	TBD
9	Ruminative Responses Scale	5	TBD
10	Temporal Discounting	10	TBD
11	Pittsburgh quality of sleep index	5	TBD
12	Positive And Negative Affect Schedule (PANAS)	5	TBD
13	Levels of optimism thinking. LOT-R (Life Orientation Test-Revised)	5	TBD
14	Risk attitude	7	TBD
15	The Interpersonal Reactivity Index	7	TBD

16	The Berkeley Expressivity Questionnaire	5	TBD
17	Stop Signal Task	20	TBD
18	NEO- Personality Inventory - Revised Test	35	TBD
20	Lottery choice experiment	5	TBD
21	Autism Social Responsiveness Scale (SRS)	20	TBD
22	The Emotion Regulation Questionnaire (ERQ)	5	TBD
23	Stroop effect test	5	TBD
24	Moral foundations questionnaire	5	TBD
25	Implicit Association Test - Race	8	TBD
26	Implicit Association Test - Gender	8	TBD

6.5 Diet

Monitoring dietary intake is an active area of technological exploration, with solutions ranging from cameras to record subjects' meals to skin sensors to detect change in subjects' chemistry. Unfortunately none of these solutions seem ready to be deployed outside the laboratory yet. Consequently, KHP study will ask a subset of the subjects in the core cohort to keep a diet diary for one week at a time at regular intervals, once or twice a year. In addition, subjects will be asked to photograph their meals and upload them for future use with machine learning applications.

During the week of keeping a diet diary, the study will also experiment with reminders and prompts to increase the recall rate of dietary information, such as sending messages to subjects to inquire whether they just had a meal or a snack, based on time of day, times of previous diary entries, location of the subject (for instance exiting a restaurant or a bodega) and other relevant factors.

In order to estimate the caloric value and nutritional content of food KHP servers will reference external databases, such as USDA's Nutrition Database.

Table 3-15: *Diet Measurements*

No	Data Element	Frequency
1	Food item(s), time of consumption	1-week interval, annually
2	Beverage item, time of consumption	1-week interval, annually
3	Picture of the food item	1-week interval, annually

6.6 Educational

KHP will gather several categories of educational data with a particular focus on children and teenagers in the study. For adults in the household, KHP's primary interest is highest education level achieved and the time of latest schooling. For children and teenagers, KHP will gather information about both formal and informal education. Formal education data will include, not only each student subject's grades, attendance and disciplinary record for each year, but also information about the school itself as maintained by New York City Department of Education (NYC DOE). Data on information education will include items such as the number of books in the home and student's participation in extracurricular activities, such as music, arts or sports.

Table 3-16: *Education Measurements*

Formal Education for Adults

No	Data Element	Frequency
1	Highest level of schooling	Once
2	Schools, degrees, majors, dates of graduation, GPAs	Once

Formal Education for Children and Teenagers

No	Data Element	Frequency
1	Student's school and grade level	Annually
2	Student's school subjects, teachers' names, class sizes and grades	Annually
3	Student's absence record	Annually
4	Student's disciplinary record	Annually
5	Standardized test results for the student	Annually
6	Teachers' report on the student in prose	Annually
7	Participation in school-related extracurricular activities	Annually
7	School report by NYC DOE	Annually

Informal Education for Children and Teenagers

No	Data Element	Frequency
1	Number of books in the home	Once (at intake)
2	Participation in activities outside the school (e.g. camps)	Annually
3	Work experience (after-school or summer jobs) – duration, hours / week, job responsibilities	Annually

6.7 Occupational

KHP will seek both an occupational history of adult subjects at intake, and will follow subjects' occupational trajectory through the remainder of the study utilizing W-2's as well as self-reported data.

Table 3-17: *Occupation Measurements*

No	Data Element	Frequency
1	List of past jobs held: dates, companies, locations, titles	Once (at intake)
2	Updates to employment status	Annually

6.8 Physical Activity and Mobility

KHP will collect data on physical activity and mobility primarily through the KHP app, which will utilize activity trackers and location tracker modules in Apple and Android smartphones. In addition, detection of Bluetooth beacons from known KHP entities either in the home or other family members will be captured.

Table 3-18: *Physical Activity and Mobility Measurements*

No	Data Element	Frequency
1	Latitude	Every 5 minutes
2	Longitude	Every 5 minutes
3	Daily distance traveled	Daily
4	Daily steps taken	Daily
5	Hourly distribution of distance traveled	Daily
6	Hourly distribution of steps taken	Daily
7	KHP Bluetooth beacon id, RSSI, time of RSSI measurement	Ongoing
8	Other Bluetooth device ID, RSSI, time of RSSI measurement	Ongoing

Note: RSSI: Received Signal Strength Indication, which is used to estimate the distance between the subject and another Bluetooth beacon or device.

6.9 Social Media and Digital Screen Time

KHP app will collect information about when subjects engage on social media apps and total screen time on the smartphone.

Table 3-19: *Digital Screen Measurements*

No	Data Element	Frequency
1	Phone screen lock / unlock times	Every 15 minutes
2	Social media app name, activity start & stop times	Every 15 minutes
3	Number of incoming and outgoing SMS	Every 15 minutes
4	Number of incoming and outgoing MMS	Every 15 minutes
5	Number of incoming and outgoing e-mails for each active e-mail account on mobile device	Every 15 minutes
6	Number of incoming and outgoing calls	Every 15 minutes

In addition, KHP servers will query publicly available APIs from social media sites to get additional information on the number of social media updates that subjects make and receive on a daily basis.

Table 3-20: *Social Media Measurements*

No	Data Element	Frequency
1	Social media app name	Daily
2	Number of contacts on the app	Daily
3	Number of incoming updates to the subject on the app	Daily
4	Number of subject's updates ("like"s, posts) by type on the app	Daily

6.10 Caregiving in the Home, Family Interactions, Social Contact

KHP will use two data sources to calculate the scope of caregiving by family members and non-family. One source will be the intake interview during which subjects will be asked to identify those family members who require care and their caregivers. In addition, subjects will be asked to provide information about how frequently they socialize with immediate and extended family members, as well as the scope of socialization with friends and colleagues. The second source of data will be collected by the KHP app to detect known Bluetooth beacons or devices that will be worn by family members, and to keep a running list of other Bluetooth devices that it detects during the course of each day.

Table 3-21: *Interaction Measurements*

No	Data Element	Frequency
1	Bluetooth ID detected, for all such beacons and devices	Every 15 minutes

6.11 Financial

Financial information covers each subject's credit rating and four other major categories: income, expenditures, assets and liabilities of each subject. Within each category KHP will capture each transaction. For income and expenditures KHP will access bank accounts, credit cards and debit cards, with subjects' consent, using an app like Mint.com, which may be built in as a capability into the KHP app. In addition, we are exploring the viability of using a data aggregator, such as Yodlee.com as a secondary source for ensure completeness and validation of this data.

Assets include all cash and non-cash assets with a value above a threshold, which is TBD. Such assets include property, vehicles, retirement accounts and investment accounts. Liabilities include mortgages and debt with a value above a threshold, which is TBD. Income sources include employment income, rent, social assistance programs and other sources. Expenses include rent, mortgage payments, car payments, insurance, utilities, food & beverages, clothing, entertainment and several other categories.

Table 3-22: *Financial Measurements*

No	Data Element	Frequency
1	Subject's credit rating by Experian, Equifax, TransUnion	Annually
2	Credit card transactions	Monthly
3	Account name and type (e.g. Savings account, credit card, mortgage, insurance)	Monthly
4	Account value at the end of the month	Monthly
5	List of payments received in account by source, date & time, and received amount	Monthly
6	List of payments made from account by payment category, date & time and payment amount	Monthly

KHP will also gather information on taxes that are paid by each subject along the following categories.

Table 3-22: *Financial Measurements (continued)*

No	Data Element	Frequency
1	Federal income tax	Annually
2	State income tax	Annually

3	City income tax	Annually
4	Employment tax	Annually
5	Social Security tax	Annually

6.12 Interactions with Law Enforcement and Justice Administration

KHP will gather information about subjects' interactions with local Law Enforcement agencies as both victims and alleged perpetrators, primarily using data that is maintained by New York City Police Department and District Attorney's Offices. NYPD maintains databases on Stop & Frisk activities, moving violations and arrests.

In addition, KHP app will capture the date and time of each 311 and 911 call that is made from a subject's phone.

Table 3-23: *Judicial Measurements*

No	**Data Element**	**Frequency**
1	Subject's interaction type with NYPD (e.g. call to 311 / 911, complaint, moving violation, arrest)	Monthly
2	Date & time of event	Monthly
3	Subject's interaction with DA's Office after an arrest, date & time of update	Monthly
4	Subject's calls to 311, dates & times	Monthly
5	Subject's calls to 911, dates & times	Monthly

6.13 Personal Interview and Diary

The KHP app will provide the ability to perform a video interview of subjects at select intervals, which will be used for the subcohort only. The video interview capability will prompt subjects with one or more questions, one question at a time, and record the audio and video response of the participant. Questions will be displayed either as text on the mobile device, or read out loud by a text-to-speech capability, i.e. the interviews will not be conducted as a live 1-on-1 interview with a researcher. Responses will be codified, initially by a human and eventually by an intelligent Natural Language Processing (NLP) system to extract and digitize key information. Subcohort subjects will be prompted for a video interview every 3 years.

In addition, the KHP app will enable and encourage all core cohort subjects to keep an audio and/or video diary using the app. The app will upload entries from the diary to a data collection server when WiFi connection is available, or if upload over cellular is enabled. In addition, participants will be able to store their diary entries on any long-term storage medium they prefer, such as iCloud, Dropbox, Google and others. The intended use of this module is to gather free speech samples, which can be analyzed for predictors of physical illness, cognitive decline or mental disorders by researchers.

6.14 Surveys

The KHP app platform will enable researchers to survey any subset of subjects in support of future research efforts, in addition to standardized gamified psychology questionnaires.

6.15 Neighborhood Baseline

Neighborhood baseline will consider the census data from American Community Survey using Zip-code Tabulation Areas (ZCTAs) along the following categories: Social, Economic, Housing and Demographic (see Table 3-24).

Table 3-24: *Neighborhood Measurements*

No	Data Element	Frequency
1	Social census data for subject's zip code	Every 3 years
2	Economic census data for subject's zip code	Every 3 years
3	Housing census data for subject's zip code	Every 3 years
4	Demographic census data for subject's zip code	Every 3 years

7. Known Issues

7.1 Known Gaps in Data Collection

KHP faces the limitations in obtaining firsthand data in the following domains:

- Cash transactions
- Barter economy
- Non-phone communications (Skype, Google-chat, etc.)
- Browser and search history on non-phone devices (home and work computers, Windows or Mac) and on 3rd party devices, e.g., library computer
- Comprehensive "Screen time" across TV, tablet, smartphone and e-Readers

7.2 Known Limitations with Statistical Power

Given resource limitations, in both funding and subjects' time, KHP also faces potential issues in certain areas due to the size of the core sample or the sampling frequency of certain data points.

While the core sample is large enough to address a large number of questions regarding the general population, its "effective size" for a particular study, which impacts only a subsection of the population (e.g. certain genetic variations), will be significantly smaller.

Similarly, we are mindful that the 3-year sampling period for some biological data points within the 10,000-person core sample may lessen the immediate value of that data until the effective population size expands via crowd-sourcing efforts for those data points or through additional studies in other cities. Two such data points are microbiome and hair cortisol levels. Gut microbiome has high variability month-to-month; consequently, a 36-month period between samples may or may not prove to be too large to leverage microbiome data immediately for research. Cortisol measurement from each hair sample only indicates the level for the previous month, which leaves a 35-month gap between data samples.

There are several possible approaches to increasing sampling density for select measurements. One possibility would be to take more frequent measurements for a limited period of time in the subcohort. Another would be to allow researchers with specific hypotheses and associated funding sources to do more detailed measurements with a subset of the main study population. We are also exploring whether data mining methods may offer solutions that would improve the utility of KHP data in the near term.

4

STUDY FRAME DESIGN

1. Study Frame

The goal of the Kavli HUMAN Project (KHP) is to examine the lives of a population-based sample of roughly 10,000 New Yorkers over time. In order to generalize the findings of the KHP to the entire population of New York City, we need to select a representative sample of New Yorkers. A representative sample is an unbiased subset of a statistical population that accurately reflects the members of the entire population. We will select this subset using a probability sampling method, which is a method in which every sample unit in the population has a known, non-zero probability of being included in the sample. The combination of these methods will make it possible to produce unbiased estimates of population parameters.

The study will require a *Master Study Frame* that will be created by combining administrative and purchased data sets, and using a multi-stage area probability design to sample and recruit approximately 2,500 New York City households. Within the sampled household, a primary participant will be selected and these primary participants will be a representative sample of New York City residents. Additionally, in order to capture unique age-specific experiences of our populations of interest, we plan to oversample three groups: young children (aged 0-3 years), pre-teenagers (aged 5-9 years), and seniors (aged 65+).

1.1 Administrative Data Sources

The sampling unit of the study frame will be residential addresses in New York City. One source of this data is the Primary Land Use Tax Lot Output database (PLUTO), which contains information ranging from community district and census tract to the number of residential and non-residential units in multi-unit buildings. Additional variables in this dataset that could be useful for other analyses in the project include: school district, city council district, police precinct, fire company, and the geographic x,y coordinates.

Each apartment building or condominium complex appears only once in the PLUTO database and the total number of units in each building is listed. A limitation of this database is that the individual level unit addresses are not included. Additionally, commercial spaces that have been illegally converted to residential occupancy and residential units that have been subdivided are not counted. There are neighborhoods in the city where this is a present concern, and thus using only PLUTO we may under-sample residential units in some areas. These limitations can be corrected in a number of ways, such as modeling using techniques developed by the Center for Urban Science + Progress (CUSP) and the use of supplemental data sets.

1.2 Potential Sources for Supplemental Data on the Oversample Groups

A proposed data source for information on the youngest potential Kavli HUMAN Project participants are the aggregate birth certificate data that can be requested from the New York City Department of Health and Mental Hygiene (NYCDOHMH), Office of Vital Statistics. Information available on the birth certificate includes: parents' borough of residence, zip code, Community District, age, race/ethnicity, educational attainment, and form of insurance.

An administrative data source that could be used for school-age children is enrollment information

collected by the New York City Board Of Education. The study will request school enrollment data by grade and home zip code or school district through the Research and Policy Support Group (RPSG).

Another potentially important source for student records is the National Student Clearinghouse. This resource contains a nearly complete database of student enrollment and degree records. It provides reporting services and regularly collaborates with institutional researchers.

Among the many potential sources of information about seniors in New York City, we propose using data gathered by the Centers for Medicare & Medicaid Services (CMS). CMS is responsible for administering the Medicare, Medicaid, and State Children's Health Insurance Programs, as well as a number of health oversight programs. CMS gathers and format data to support the agency's operations. Information about Medicare beneficiaries, Medicare claims, Medicare providers, clinical data, and Medicaid eligibility and claims are included. The CMS has three levels of data specificity and privacy review required for release of data to researchers: identifiable data files contain actual beneficiary-specific and physician-specific information; limited data set files contain beneficiary level health information, but exclude specified direct identifiers as outlined in the HIPAA Privacy Rule; and non-identifiable data files contain non-identifiable person-specific information and are within the public domain. Each of the three categories has its own process for requesting data.

A limitation common to administrative data sources is the finest level of geographic detail is often zip code, in order to prevent individuals from being identified. However, a strength of using this level of data is that zip code is one variable that is nearly always present in administrative and survey databases, and these data sources can be combined to build up a fuller picture of the population in a specific geographic area. Over the course of the KHP we will be using data that have differing levels of geographic granularity, and we will build on the established data science techniques to properly merge these data sets.

1.3 Purchased Data Sources

Data sets are available to be purchased from commercial vendors serving academic, survey, and market research organizations. The cost per record is dependent upon the type and specificity of information requested.

1.3.1 Address-Based Sample (ABS)

The address of residence is the basic sampling unit and we propose using a targeted address-based sample generated by the US Postal Service Computerized Delivery Sequence File (CDSF). This data source provides complete and accurate addresses with the ability to identify addresses marked as vacant or seasonal. Each record is a complete mailing address and specifies the type of service delivery, and the CDSF is updated monthly by the Postal Service. In its raw form the CDSF is simply a database for delivery of mail and does not contain any information about the composition of the household, but a commercial vendor can add this information. Additionally, because there is not a one-to-one correspondence between the USPS and US Census geographic definitions, additional information about the addresses needs to be appended in order for Census geographic definitions to be used for sampling. Geo-coding each address to a unique Census block can bridge the gap between USPS and Census data. While accommodating the geographic needs of developing a sampling frame, this enhancement also allows appendage of many ancillary data items to each address, including those available from the Census and commercial sources.

1.3.2 Listed Household Sample (LHS)

Listed Household Sample is a telephone sample pulled from a number of different commercial databases that individual marketing vendors license. LHS can be further targeted by a number of household demographics including age, gender, race/ethnicity, and income of the head of household. Additional flags can be appended such as for households that are likely to include children under

18. Telephone numbers are limited to those published in white page directories, but a cell-phone append can also be done where available.

1.4 Summary of the Sample Frame Composition

Administrative data sources are often consulted first to develop a sample frame because they are readily available to the public and contain the information needed to define a population or geographic location. However, production and reporting schedules vary greatly with these sources, so they may contain out of date information and can lack the level of specificity needed for sampling.

Purchased databases are very often used in survey research due to the ease of use and the information contained being updated continually, but they are not without limitations. A drawback of using the postal service CDSF as part of a sampling frame is more apparent in rural areas, where mail delivery may be to a post office box or general delivery and not to an actual physical location. Additionally, some households will have mail delivery to both a residential address and a P.O. Box, which could lead to frame multiplicity as such households would have multiple chance of selection. We do not anticipate this to be a problem in New York City because we will be using only residential addresses, not P.O. Boxes in the sample frame. However we acknowledge that by using residential addresses as the sampling unit we will automatically be excluding people from our study population who are homeless or are in housing facilities (military, universities, hospitals, prisons) at the time of enrollment and we will statistically correct for these exclusions.

We think that starting with a purchased address-based list sample and supplementing it with Census and commercial geographic and demographic variables will build the best sample frame for the HUMAN Project. The postal service CDSF also identifies those addresses that are commercial, so that these can be removed before the sample is drawn. The PLUTO database could be added as a layer of tax and property information needed for the study, but that database is not needed to draw the sample frame. The purchased address list should cover all potential units for sample selection without frame multiplicity.

The geographic unit we will use to construct the Master Study Frame is census tract, and we will use this variable to map all of the data gathered from various sources into the model. This model will be dynamic over the lifetime of the study as addresses are added and removed from the postal service CDSF, and other possible geographic or demographic features become available with purchased list samples. Additionally, illegal conversions of commercial spaces into residential spaces and subdivision of existing residential space are factors common to New York City housing, and are not captured in administrative or commercial databases. CUSP used data on electricity, water, and sewage use to develop estimates on the fraction of illegal conversions within each census tract. We will work with information developed by CUSP on methods to estimate the extent of illegal conversions in the Master Study Frame.

Working with the Study Frame Advisory Council, the Study Population Officer and the HUMAN Project Chief Demographer (to be hired) will develop and maintain the Master Study Frame. The frame will be updated regularly, ensuring that throughout the course of the project we will have the most accurate sample of New York City households.

2. Building a Study Population

Our goal is to recruit approximately 2,500 study households throughout the five boroughs of New York City based on the Master Study Frame (Figure 4-1.) We propose using a multi-stage area probability sample design that will minimize sampling, coverage and non-response error, while providing a statistically representative sample of New York City residents.

2.1 Residential Unit Selection

Using the multi-stage area probability sample design method, which is based on the sampling method used for the U.S. Health and Retirement Study, there would be four selection stages: (1) selecting probability proportionate to size census tract segments across the five boroughs; (2) sampling of area segments within the sampled primary census tract segments; (3) systematic selection of housing units from the sampled area segment; and (4) selection of a "seed" participant within the household.

2.2 Participant Selection

The standard model for study participant selection will be modified for this study. In consultation with the Study Frame Advisory Board and our statistical experts, we will develop a sampling algorithm based on the number and ages of people in the household. Within the sample households, we will select individuals or "seeds" and then build the study population outwards from them, similar to the participant and spouse selection in the Health and Retirement Study. The resulting distribution of the age of "seeds" will be the sum of the inverse probabilities of selection among residents of all agesin NYC. This will not match completely the age distribution of NYC residents, because we will be oversampling three age groups (0-3 years, 5-9 years,

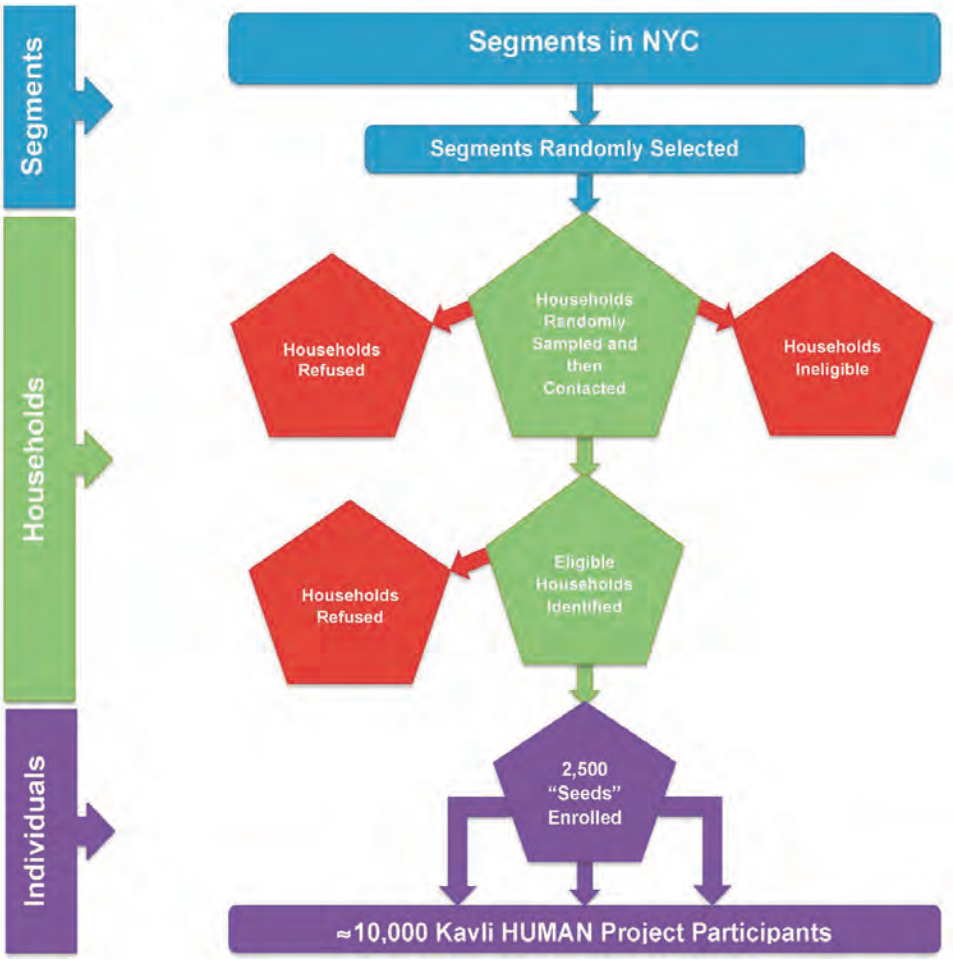


Figure 4-1: *Kavli HUMAN Project Sample Design*

and over 65). We expect that we would need to successfully recruit about 2,500 seeds in order to fill the study with 10,000 people.

2.3 Inclusion Criteria

There will be two criterion that must be met in order to be a participant in the HUMAN Project: the first is the ability to communicate in English at a level to be determined in CD2, and the second is to be a resident of the selected household.

2.3.1 Language Considerations

New York is a multi-lingual city, with an estimated 200 languages spoken. However, four languages account for an estimated 84% of the languages spoken in the city. In descending order, the top four languages spoken by New York City residents are English, Spanish, Chinese, and Russian. Ideally we would like to capture the experiences of New Yorkers who communicate in languages other than English or Spanish. We recognize the research limitations of using just two languages, yet the additional cost of developing, testing, implementing, and staffing the study in other languages would be an enormous expenditure in order to capture a small percentage of respondents. Using data from the 2007-2011 American Community Survey, below are three options we considered with regard to the study languages.

Option 1a: English only is spoken by 51.5% of New York City residents, and 24.6% of the city population speaks Spanish. Among the New York City population, for those who speak another language at home, 12.2% state that they speak English "very well". If we conduct the HUMAN Project in English and Spanish only, and include in the study population non-native English speakers who can speak and understand Standard English "very well", then we estimate that we will cover 88.3% of households in the city.

Option 1b: Among the New York City population, for those who speak another language at home, 6.2% state that they speak English "well". For these respondents all the study-related material will be in "Simple English" which is a subset of Standard English that uses a limited vocabulary, short sentences, and simplified grammar. We estimate that by conducting the HUMAN Project in Standard English, Simple English, and Spanish, we will cover 94.5% of households in the city.

Option 2: Chinese is next largest language group in the city with 5.5% of NYC residents speaking this language at home. Among the New York City population, for those who speak Chinese at home, 3.7% speak English less than "very well." If we conducted the HUMAN Project in Standard English, Spanish and Chinese, we estimate that we would cover 92% of households in the city.

Option 3: Russian is spoken by 2.4% of NYC residents. Among the New York City population, for those who speak Russian at home, 1.5% speak English less than 'very well." If we conducted the HUMAN Project in Standard English, Spanish, Chinese and Russian we estimate that we would cover 93.5% of households in the city.

At CD2 we will construct a cost/benefit analysis to examine the inclusion of each additional non-English/Spanish speaking respondent. Preliminary plans lean toward options 1a or 1b.

2.3.2 "Seed" Respondent Characteristics

The seed HUMAN study participant must be a permanent resident of the sampled household and spend at least 50% of their time in that residence; must speak and understand English or Spanish (possibly others); and must be able to give informed consent personally, or by parental proxy for those under age 18. For those older than 65, and as needed with other potential participants, we will conduct a short standardized test to screen for cognitive deficits that would indicate a proxy respondent will be needed in order for the selected respondent to participate in the study.

If the selected seed participant is disqualified due to language limitations, or if a household member is

unavailable or unwilling to be a proxy respondent for the individual with cognitive deficits, the household will be screened out as ineligible.

2.4 Residential Network Group

The HUMAN Project is not a family-based study, but to increase understanding about the influence of family and residential co-habitants on the lives and biology of our participants, particularly when those participants are children and seniors, we propose building small networks, basically residential families, around our seeds and include all of those residential family members in the study cohort. Family members, defined by either biological or legal relationships, are the people who are most likely to contribute substantial long-term influences on our seeds, and familial data substantially increases the power of genetic analyses.

We call this group a "residential network group." We define co-residence as living together at least 5% (maybe more) of the time, and we will use the US legal federal definition of family, which is "any individual related by blood or affinity whose close association with [target individual] is the equivalent of a family relationship." We are considering the consequences to the sample representation of requiring agreement from all members of the seed's residential group to participate in the study in order for the seed or any of the members of the group to be enrolled. If we require agreement from every member of the residential network group, we anticipate that will have negative consequences for our cooperation rate and the composition of the study sample. Conversely, if we do not require all members of a residential network to participate, we will have an incomplete picture of the seed's environment and experience. At CD2 we will examine further the implications of both courses of action.

Some residential units will separate after enrollment, and we have considered two options for following the seed and the residential network group. In the first option, we continue to follow the seed, and every residential network group where the seed lives at least 50% of the time. In some cases this definition of a residential network group would result in a rather large residential network: for example, if the seed is a child of divorced parents with joint custody, the network for this seed that included his/her biological parents, as well as any step-parents, step-siblings, or half-siblings who live with him/her. The second option is to restrict the definition of the seed's residential network group to those people in the household at the address where the seed was recruited. We will work with the Study Frame Advisory Council to further explore the implications of these two options, and during CD2/3 we will decide on a strategy for following the seed participant in the event of household dissolution.

Roommates who do not meet the federal definition of family members will not be enrolled in the study when a seed is identified. Adult children of a seed who live at college during the academic year would be enrolled as study subjects if they met the minimum residential time requirement.

According to the American Community Survey 2007-11, approximately half of all Manhattan residents live alone and across the city the percent of single resident households is around 33%. Additionally, nearly 70% of households have no related children under 18 years old living in the unit. These two features of the New York City household composition suggest that we may need to recruit more than the estimated 2,500 households in order to reach the goal of 10,000 HUMAN Project respondents. Utilizing the household level data appended to the purchased address list, in CD2 we will develop sampling strategies for optimizing the number of study participants we could potentially get from each selected household. We will keep the seed participant sample balanced to be representative of the New York City population, while developing a sampling plan to yield the greatest number of study participants from each sampled household. We will study the costs and benefits of this type of sampling approach and balance that against increasing the goal of 2,500 households to account for the high number of single

resident and child-free households in New York City.

2.5 Non-residential Network Group

In order to get the most complete picture possible about our seed participants, we propose to gather genetic information from their non-residential family members. The level of data collection from these individuals will not be as deep as the residential network group but it will add important information about the environment surrounding the seed, such as the influence of non-custodial parents, or adult children who are the care takers for parents with whom they do not reside.

2.6 Nesting Samples

To address the twin limitations of cost and scale, we propose to nest the core sample group of 10,000 HUMAN Project participants within larger and smaller subgroups (Figure 4-2.)

2.6.1 Intensive Study Sub-cohort

From the primary cohort of 10,000 residential unit participants, we will recruit a statistically random sample of 250 individuals who will participate in more detailed measurements. These participants will be invited to complete higher cost biological tests such as brain scans and extensive genetic testing. These tests will either be performed by the study itself or by groups collaborating with the study. The goal is to use the highest density sub-sample as a calibrating tool for a deeper understanding of as large a population as possible.

We will develop the sampling, recruitment, and retention policies for the sub-cohort of 250 respondents in CD2/3. These individuals will be participating in the sub-cohort data collection and biological testing in addition to the ongoing data and biological collection efforts in the core population. In later documents we will specify the types and frequency of data to be collected for these participants.

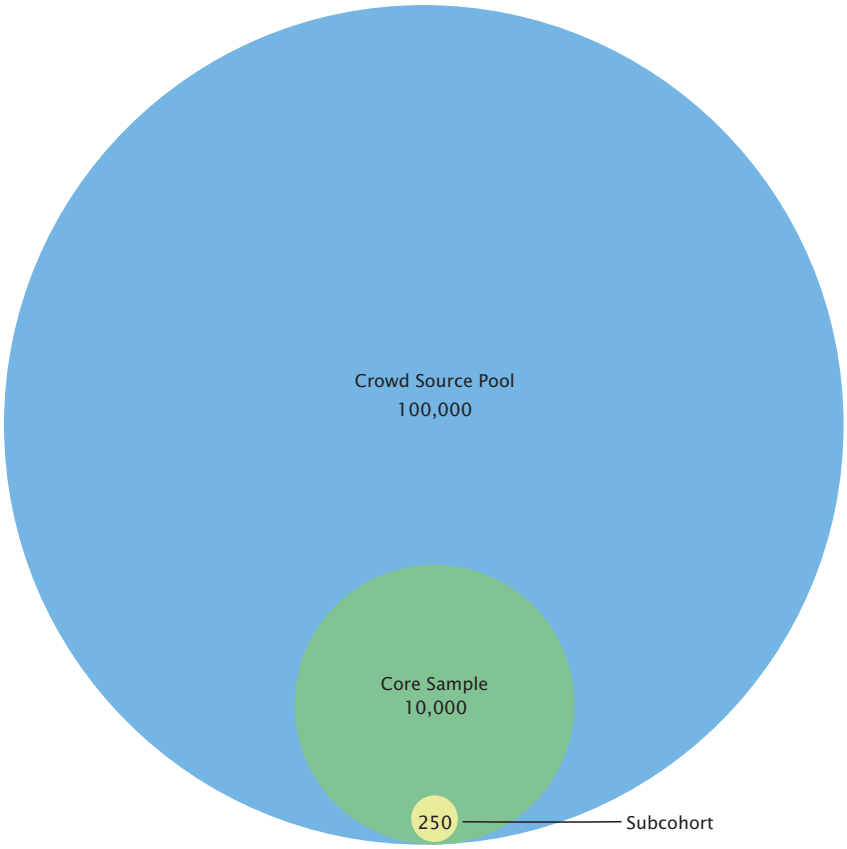


Figure 4-2: *Nesting of Kavli HUMAN Project Samples and Number of Participants*

2.6.2 Extensive Crowd-source Cohort

We suggest a public outreach effort that leverages web and smartphone media to engage a large group of individuals interested but not chosen to participate in the HUMAN Project. This will engage those people most interested in this type of "quantified" living, and it will give the HUMAN Project another data source and outreach vehicle to contact with New Yorkers.

We propose asking these people to download the Kavli HUMAN Project mobile application and participate in many of the Project tasks. Their data will be kept separate from the core sample population and they will receive general feedback on the crowd-source population metrics on the KHP website. Our goal will be to use this data efficiently by calibrating it against the detailed high reliability 10,000 person core population.

The timing for enabling enrollment in the crowd-source cohort will be further developed in CD2/3. Ideally, we would like to have the crowd-source structure in place at the time of core sample recruitment, so that the data are gathered are across the same time frame and to allow people who are immediately interested a chance to participate in the study.

3. Pre-recruitment

A critical component of recruitment for the study will be the "pre-recruitment" phase in which we propose a series of community gatherings, information sessions and focus groups to discuss all aspects of the project, to elicit questions, and to address all questions and concerns. It is particularly important that we spend time in minority and under-served areas in order to develop trust among the residents, who are generally less willing to participate in research studies. We will coordinate with and use the expertise of the Education and Public Outreach group to develop these community activities.

3.1 Community Outreach

Once the formal HUMAN Project advertisement campaign has begun across the city, we propose a number of different approaches to reaching out to smaller segments of the population to disseminate study information and answer questions. We will coordinate these activities with the HUMAN Project Education and Public Outreach group.

There are 59 Community Districts in New York City. One strategy we propose is contacting a sample of active Community Boards in each of the boroughs to speak about the study at open sessions and seek their endorsement.

We will also reach out to knowledgeable community leaders or organizations to request time to speak or present information about the study at neighborhood-specific events.

Parent organizations could be a great opportunity to engage people and to speak extensively about the role children would play in the study. During the school year, we propose selecting a small number of active parent organizations to contact and request time to speak at one of their meetings.

3.1.1 Kavli HUMAN Project Advance Communication

We propose hiring undergraduate students enrolled at NYU and in the City University of New York (CUNY) system to form a pool of part-time temporary workers for neighborhood advance teams. These teams will go into neighborhoods shortly before the recruitment teams start their in-person information sessions. These advance teams will get information about the HUMAN Project out at the street level by attending neighborhood events such as those listed below, and through venues which have proven successful in other population-based studies, such as churches, barber shops, and beauty salons as appropriate.

Street fairs are held every weekend in New York City, from roughly Memorial Day through Labor

Day. We could rent and staff a booth for at least one street fair within each borough to answer questions, get the logo out in the community, and distribute pamphlets and study embossed merchandise.

Farmers markets are another popular mobile resource we could use to connect with residents across the city. These markets tend to operate from early spring through the fall, and most accept supplemental nutritional assistance cards, so they are utilized by some of the under-served populations we want to educate about the study.
We will be engaging the potential study population using a variety of in-person methods to provide information about the project and we will also be using paid media communication that is geographically targeted (See the Education and Public Outreach chapter.) As part of the final study frame sampling plan we will have a schedule of the neighborhoods where the recruitment teams will be each week over the course of the project. We will conduct a coordinated, geographically targeted information campaign that will have the effect of "priming" the neighborhood residents with positive, credible information about the project ahead of the initial visit by the enrollment team. The in-person effort will be supplemented by an online, direct mail, and print information campaign. Four to five weeks before a recruitment team is scheduled to be in a neighborhood, residents of the sampled households will receive the HUMAN Project alert postcard (see section 5.1.) One to three weeks before the recruitment teams arrive in the neighborhood, residents will start to see information about the HUMAN Project in the local newspapers, in other print media, and online. The selected households will also receive the HUMAN Project information package. As contact is made with the selected households, further methods of paid communication can be specifically targeted to those residents and fill the information gaps between visits from the recruitment team with persuasive, positive messages.

3.2 Qualitative Research

In July 2015, David Kaufman, Program Director at National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) gave a talk at the Precision Medicine Initiative (PMI) Participant Engagement and Health Equity Workshop. The workshop was on participant engagement and health equity as they relate to the proposed PMI national research cohort. The talk was an overview of results from a recent national survey on public attitudes about the PMI Cohort Study that was sponsored by the NIH Common Fund. This survey asked how people feel generally about the idea of the PMI cohort, if they would be willing to participate in this sort of cohort, and feelings about the cohort itself. The survey found that 54% of respondents said that they definitely would or probably would be willing to participate, and willingness was fairly uniform across demographics. A detailed summary of this survey can be found in Appendix I.

The results of the public attitude survey regarding the Precision Medicine Initiative are informative for the HUMAN Project. We are encouraged by the general acceptance of the concept of such an in-depth study of health and behavior, and we see the need to conduct New York City-specific qualitative research. We will work closely with the HUMAN Project Education and Public Outreach group to conduct the preliminary research to get feed-back from people on the study in general, concerns or reservations about types of data we want to collect, how that data will be used, and any considerations for special populations such as children and seniors.

3.2.1 Key Informant Interviews

We will conduct key informant interviews at the individual level, and with people who are within organizations and institutions that will be able to contribute valuable information to the planning of the HUMAN Project and may have an interest in the outcome.

We plan to identify 20-30 members of the community across the broad spectrum of socio-economic levels within New York City who understand the need and potential benefit of the type of data the HUMAN Project will generate, in addition to those who are living or working in communities where we anticipate more of a challenge with recruitment. These individuals will be able to speak to questions such as: what are your initial responses/reactions when presented with information about the project? What are the biggest anticipated barriers to participation? What are the attitudes toward financial compensation for participating in the study? What do you think are the issues that would be most important to the populations (parents, seniors, cultural communities, religious groups, non-native English speakers, and undocumented immigrants) that you serve? What are some of the issues facing these sub-populations that we have not identified? The information we receive from these interviews will help drive the content of the focus group guides.

A number of longitudinal health studies have been conducted in New York City and it will be beneficial to learn about these experiences. RTI, a survey research company based in North Carolina, conducted the New York City Health and Nutrition Examination (NYC-HANES) in conjunction with the CUNY School of Public Health for the Department of Health and Mental Hygiene in 2013. Many of the sampling and data collection modalities that were used in the NYC-HANES project are similar to those proposed for the HUMAN Project. We believe that it could be helpful to speak with project leadership from RTI, CUNY, or NYCDOHMH familiar with the NYC-HANES project to get at best practices and lessons learned from that experience as we finalize our strategies.

In addition to the experience with NYC-HANES, we believe that speaking with leadership at the NYCDOHMH will be beneficial for the success of the project. NYCDOHMH conducts many on-going research projects in the city, and we think that we should seek to gain insight into their experience with population-based research with New York City residents. Additionally, NYCDOHMH buy-in and support of the project would be beneficial as they have a presence in many of the under-served communities where we anticipate we may have difficulty with recruitment, and any comment, either positive or negative, from this source may have an impact on our study.

We want to involve New York City and State government in the HUMAN Project. We will work closely with the Education and Public Outreach group to meet with local and state lawmakers, such as the city council, borough leadership, and state legislature, to present the study and gather input and determine areas of mutual benefit.

3.2.2 Focus Groups

Building on the information generated in the key informant interviews, we will conduct a series of 8-10 focus groups to explore these themes. Additionally, we will use these opportunities to finalize messaging strategy and the execution of recruitment messaging and materials. We will conduct focus groups in English and Spanish, and at least one session will be held in each borough.

We propose conducting sessions with adults where we would utilize the information learned during the key informant interviews to outline the structure of the study, and go into detail about the type of data we plan to collect, and the method of collection. We want to delve deeply into topics such as: how many visits by the recruitment team would be necessary or what is the level of information that the potential participant needs to understand before we request that a household enroll in the study? What kinds of bio-specimens would people be most willing and most reluctant to provide? What are the sources or types of data that people consider too personal to share? What role, if any, do incentives play in influencing these decisions? What types of incentives (monetary, information, products, services) do people think would be most important in this kind of study?

Additionally, we may conduct a focus group session with individuals who self-identify as technologically

challenged or who have had limited experience with the technology we plan to use in the study. We would work with the HUMAN Project Measurement and Technology group and have smartphones loaded with a prototype of the Kavli HUMAN Project mobile application (see Measurement and Technology chapter) to direct and moderate a user experience.

We propose conducting parent-only focus groups because we will be oversampling children aged 0-3 and 5-9 and anticipate that it may be difficult to recruit young children into the study. It is critical that we enroll children both as the seed study participants and as members of residential network groups, so in the session we want to explore and expand on the ideas raised in the key informant interviews regarding the causes of parental apprehension to help us to better address these concerns.

Similarly, we anticipate that we may face added challenges to enrolling seniors, which is especially important as we also plan to oversample those aged 65 and older. We will conduct a session with seniors to explore attitudes, challenges, and concerns specific to this demographic. The idea of senior enrollment is another topic that can be raised in the general adult sessions, and we may face challenges from adult children acting as "gatekeepers" for their parents.

We will analyze all the focus group data and see if we can identify any general themes and possible trends by age, gender, educational level, or socio-economic status. We will then use this information to develop strategies in CD2/3 to address these concerns in the HUMAN Project study information and recruitment materials.

4. Recruitment and Enrollment Personnel

The most important challenge of the HUMAN Project is participant recruitment, because in order not to introduce selection bias into the study, we need as many of the selected households to participate as possible. Therefore, recruitment will be a paramount function of the study, one that will require the greatest initial resources.

To develop the HUMAN Project field recruitment and project implementation teams, we would build up the existing project staff and rely on outside expertise when needed. As the HUMAN Project continues to expand, it would be an institutional asset to have the field recruitment and enrollment study teams in-house, as opposed to having this work sub-contracted to a survey research company.

4.1 Develop HUMAN Project Capabilities

The Head of Recruitment would be a newly created position. This person would report to the Study Population Officer and would be responsible for the management, scheduling, and oversight of the field staff.

Recruitment is a critical component to the success of the study. We think that given the crucial role of the in-person recruiters, who we are calling "recruitment specialists," people should be hired by the HUMAN Project staff and be trained at the Institute for Social Research (ISR) at the University of Michigan. One obstacle we anticipate in the NYC environment is the inability to directly access the front door of selected units due to security and privacy measures such as door men, locked lobby doors, and other impediments. The ISR staff has much experience with this type of in-person recruitment and interviewing and the training they provide will be invaluable to our recruitment efforts. These recruiters will form the base of a highly trained field operations staff.

Technology specialists will be hired and trained by the HUMAN Project Measurement and Technology group to the specifics of the devices and applications that will be used in this study.

Bio-specimen collection technicians and health aides could be recruited from specialized medical employment agencies. These individuals will be trained to the specimen collection protocols of the HUMAN Project.

It is not assumed that many of these new HUMAN Project members will have previous research experience, so basic principles of conducting research will be a necessary part of their training, in addition to the specific protocols of this study.

4.2 Composition of HUMAN Project Field Teams

4.2.1 Recruitment Teams

We will have sent introductory study materials to the selected household weeks prior to the first attempt at in person contact. Yet, this will most likely be the first face-to-face contact the potential participant has with any study personnel, and a misstep at this point could negatively impact recruitment. We propose that "recruitment specialists," do this critical and specialized job. These people will be hired by the HUMAN Project and trained at the ISR to conduct participant recruitment. ISR conducts the U.S. Health and Retirement Survey (HRS), which utilizes a similar in-person recruitment approach, and the HRS has had consistently high response rates over the course of many age cohorts. We know that the training provided by ISR will be to the highest professional standards.

To be a recruitment specialist requires an individual who is professional, knowledgeable, and personable, someone who is able to be the "face" of the study, to explain the study thoroughly in easy to understand language, and finally the ability to encourage the selected household to participate in the study. We propose that recruitment teams of two individuals, ideally one male and one female, be deployed to neighborhoods to make the first in-person contact with the households. The teams will visit their assigned neighborhood, go to each selected household, and conduct the initial 15-minute introductory session. At the end of this session, the recruiters will schedule an appointment for the next recruitment/information session. We will be conducting focus groups to gauge the number of visits or level of personal contact households need from the recruitment team in order to make a fully informed decision regarding research participation before we formally ask the household to enroll in the study. The household will be given a monetary incentive for permitting the recruiters to conduct the introductory session, regardless of whether they make a follow-up appointment with the recruitment team or not.

Over the course of recruitment, we will identify those people who are especially good at recruitment and use those individuals as "super recruiters," as is done by the U.S. HRS. These super recruiters will return to households that declined the initial in-person introductory session and those households that listened to the 15-minute introductory session but then declined to schedule the enrollment appointment. The super recruiters will attempt one refusal conversion with these households.

4.2.2 Enrollment Teams

To transition from recruitment to enrollment while maintaining good will and building trust with the household, we suggest that the recruitment specialist who has been working with the household accompany the enrollment specialist on the first visit to introduce the enrollment team to the household. The enrollment specialist will be the person responsible for documenting that all baseline data collection measures are completed.

We envision that the enrollment teams will be in the neighborhoods in specially out-fitted HUMAN Project vans. The driver of the van will be a member of the enrollment team. These vans will contain the household technology kits, bio-specimen collection kits, and the survey and psychological testing materials that will be needed for baseline data collection. The team leader, who is the in-field team manager and scheduler, will make sure that the van is properly stocked each day. Each van will be in a different neighborhood each week.

We propose having three HUMAN Project field vans, with each van having 2 assigned enrollment teams. Each in-home enrollment team will be

comprised of two permanent and three shared individuals: an enrollment specialist and a health aide will comprise one permanent team; and the two enrollment teams will share a bio-specimen collection technician, a technology expert, and a psychologist.

The health aide will be certified to perform basic measures, such as height, weight, waist circumference, and blood pressure. The aide may also assist the bio- specimen collection technician as needed.

The bio-specimen collection specialist needs to be a skilled and certified health care technician, who has experience with blood collection, preferably with pediatric and/or senior populations. Additional training for the specific bio-specimen samples (hair, urine, etc.) will be conducted once the types of samples and the collection protocols have been finalized.

The technology expert will set up the mobile data collection devices, teach the participants how to use the equipment, and fit the seed and residential group with the wearable devises. In the event that a participant is resistant to or inexperienced with new technologies, this team member will work with that individual to help them acquire the level of comfort to use the selected devise, or will recommend that an alternate device be used, such as a Bluetooth location beacon.

The psychologist is the final member of the team. This person will be a Ph.D. level clinician with the certification to administer psychological testing material.

4.2.3 Requirements for HUMAN Project Field Personnel

In order to uphold high security standards, we propose that all study personnel who will interact with participants in person be fingerprinted and complete a basic background check. At least one member of the team must be fluent in the language of the sampled household.

5. Recruitment and Enrollment

Recruitment and enrollment are two linked processes that will be conducted across a number of visits to the selected households (Figure 4-3). We believe that it is essential that the households be fully informed and given plenty of opportunity to ask questions about the study, and to also be comfortable with the study field personnel before we ask them to participate in the HUMAN Project. We will be developing the protocol for the number of visits to be attempted at each stage in recruitment and enrollment during CD2.

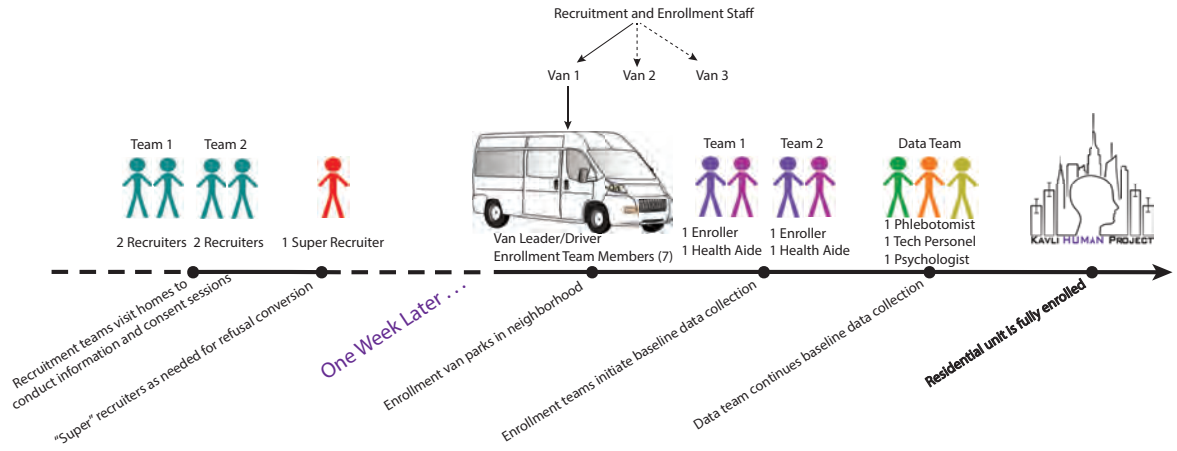


Figure 4-3: *Kavli HUMAN Project Field Teams*

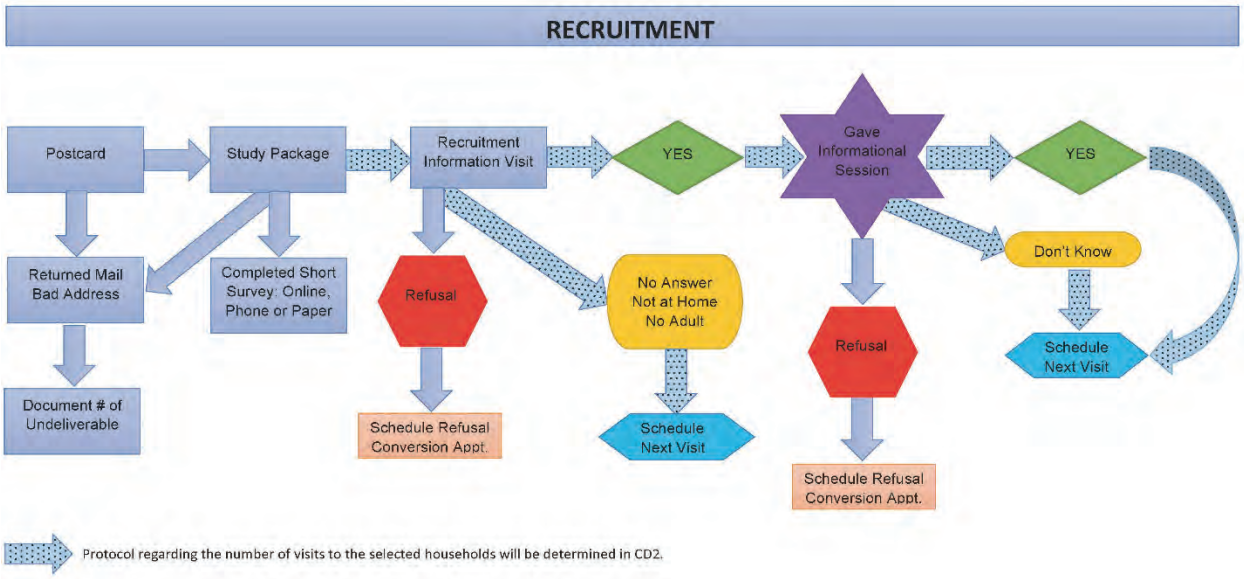


Figure 4-4: *Kavli HUMAN Project Recruitment Flowchart*

5.1 Contact with Sampled Households

The first proposed contact with the sampled household will be a HUMAN Project study alert postcard mailed approximately 4-5 weeks before the project launch. The postcard would contain general information about the project, the project website address, and a statement that they will be receiving a study information package in the mail in the coming weeks. We would recommend not including project-specific contact information for study personnel at this point to prevent households from "opting-out" before they have received the study information package.

The second proposed method of communication with the sampled household will be the HUMAN Project information package. This would contain a very detailed cover letter, explaining the study objectives, how the household was selected, that their selection is unique, and contact information for study representatives including a study designated phone number and email address. We will also include in the packet a short questionnaire. We will provide information on how the questionnaire can be completed on the web, by phone, or by returning the hard copy in the enclosed business reply envelope. We will also include a small cash incentive, maybe $2 to $5, as a good will gesture. Other items that could be included are: study brochure, FAQ sheet, and study embossed merchandise. In the cover letter we would close with a statement that field staff would be in their neighborhood within 1-2 weeks to conduct in-person recruitment. We may want to provide a way for the sampled household to schedule their consent and enrollment visits.

5.1.2 Recruitment Information Visits

The recruitment teams will go to their assigned neighborhoods and methodically approach each of the selected households (Figure 4-4). They will conduct a 15-minute information visit to introduce the project to the household, answer any questions, and enumerate the members in the household. All households that permit the recruitment team to enter and give the 15-minute information session will receive a monetary incentive. Our current plan is to have the recruitment teams visit the households (the number of times is to be determined) and once a household has agreed to participate, administer the informed consent process. We will be conducting focus groups to gain insight into how many visits by the recruitment team should be attempted before a household is asked to participate in the study. At

CD2 we will specify the number of attempts that will be made to each household and will define the dispositions to be used to code the outcome of each visit.

5.2 Enrollment

We are considering various strategies to identify the best way to conduct the consent and enrollment process (Figure 4-5). From an operational standpoint, we think that for single resident households one visit would be ideal to maximize the time spent by completing study-related consent documentation and conducting all baseline measures at one time. For larger households, it may be necessary to split the study consent process and the baseline measures into separate visits.

5.2.1 The Consent Process

The first step of enrollment is the consent process, and it will start with the study household watching a video that details the study and covers all sections of the written informed consent document that they will have in front of them. At the conclusion of each segment within the consent document, the recruitment specialist will pause the video and administer a short verbal assessment to confirm that the consent process is being fully understood by all members of the residential unit. After the video, the recruitment specialist will discuss any remaining questions and then walk the participants through signing the consent form. A copy of the signed consent form will be left with each participant.
The consent process is the action upon which all other components are dependent. Once the consent form has been signed, the order of the other enrollment tasks can be rotated as needed to get as many of the household members engaged in the baseline data collection process at once, or the other baseline measures can be scheduled for other days. However, in order to keep the chronology of data capture consistent, the time frame for completion of all measures must be short.

5.2.2 Physical Measurements

The health aide will do anthropomorphic measurements and will bring with them a stadiometer, scale, tape measure, and blood pressure cuff. They are responsible for taking the height, weight, waist circumference, and blood pressure of all the study participants, and documenting this information in the household record using a HUMAN Project laptop.

To date, we have decided to collect blood, saliva, hair, stool, and urine samples. The bio-specimen

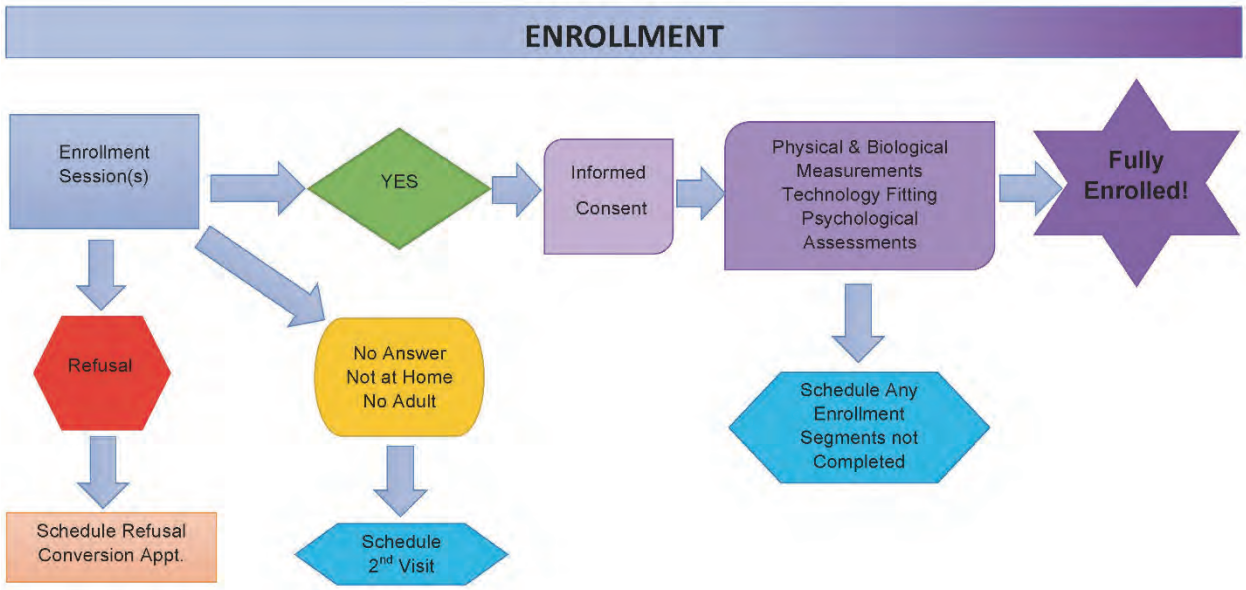


Figure 4-5: *Kavli HUMAN Project Enrollment Flowchart*

technician will have with him/her a bio-specimen collection kit that will contain all the labels and medical supplies needed to collect the samples from every residential group member. Once all the samples have been collected, the technician will return to the van to prepare and store the samples.

5.2.3 Technology Fitting and Instruction

The technology specialist will install the base monitor in a central location in the apartment. This unit will be used to sample air quality and track the respondents as they move within the household. The specialist will then work with each member of the residential group to outfit them with the technology specific to their age. Children older than 10 and all adults will be given a smartphone (if they do not already have one) that is loaded with the Kavli HUMAN Project mobile application. On either the study assigned or a personal phone, the technology specialist will demonstrate how the app works and make sure the individual is comfortable using the technology. We are considering many different wearable technologies for children younger than 10 years, including Bluetooth beacons worn on the ankle and clothing with the beacon embedded that can withstand multiple washings.

5.2.4 Psychological Testing

The psychologist will administer the interviewer-administered study instruments, and will set up any self-administered data collection modules on a HUMAN Project laptop. For the interviewer-administered instruments, the psychologist will ask the residents to select an area of the apartment that is private where the psychologist and the study participant will not be interrupted or overheard. We are developing the list of measurements to be included by age level.

5.3 Privacy Procedures

We need to consider that we will be in homes of various sizes, and must develop procedures that will balance the need for privacy while conducting the measurements and assessments with the safety of the enrollment team. For example, we may want to recommend that there must be both male and female team members in a household, that a minimum of two team members be in a household at any time, and develop policies about interacting with children younger than 18.

Another consideration is equipping one of the project vans with a participant enrollment office. This office could be used when enrolling large households to facilitate more members engaged in the baseline data capture at one time. Additionally, if potential participants state that they are reluctant to have the enrollment and data teams spend an extended period of time in their residence, the enrollment office in the van could be suggested as a place where most, but not all, baseline procedures can take place.

6. Incentive Structure

We plan to build on the years of positive research results by employing principles of the Dillman Total Design Method for the project recruitment and participant incentive structure. The Dillman method, historically used for mail and telephone surveys, is based on a tailored design that is a customized to the study population, the topic of the study, the burden on respondents, as well as the budget and length of time available to conduct the study. This method uses the social exchange theory that places an emphasis on improving trust in the legitimacy of the study, and develops procedures that create participant trust and perceptions of increased rewards and reduced costs for being a study respondent. A main feature of the Dillman Method is the use of larger incentives at the beginning of the study, and tapering down the amounts given over the duration of the study as people become more personally invested.

6.1 Three Types of Incentives

We will use three types of incentives across the span of the study: money and items that have a real monetary value, items and activities that build community good will, and giving participants

personal and population-level feedback and data. The literature shows that providing even a small monetary incentive (rather than none at all) creates an increase in response rates. Results from other longitudinal studies suggest that incentives other than money engender good will and feelings of trust with respondents, as people who feel part of something are more likely to continue participation, and over time the financial cost of these items is often less than would be spent for monetary incentives.

(1) Monetary incentives – we need to use this at the first encounter, where we have not established trust in the project or the staff. However, the use of cash or its electronic equivalent must be limited. For the low-income households, we do not want to give a cash incentive that could be construed as coercive or that may put them over the threshold to receive social services. On the other end of the spectrum, the amount of cash it may take to incentivize higher income groups may be out of range for the budget.

- The first cash payment will be the same for all households that agree to let the recruitment team into the house and listen to the 15-minute information session
- Another cash payment will be at enrollment when each household receives the same amount of money, and individuals select an age-appropriate gift from KHP branded merchandise
- A limited number of small cash payments will be made for completing specific activities

(2) We will work to build community good will throughout the course of the study and one method will be to use KHP-provided gifts and activities as incentives. These will be given for completing a task or complying with an instruction, but also used in conjunction with running of the study.

- KHP birthday and holiday cards will be mailed to each enrolled participant
- KHP newsletters will be mailed or emailed to each participant household
- KHP branded merchandise, such as t-shirts with the message "I'm a Kavli HUMAN", mugs, backpacks, smartphone cases, portable phone chargers, etc. will be given
- Bring participants together, perhaps for a forum like the World Science Festival, for a session on the study
- Set up lotteries similar to what is done at Nielsen, the media research company, where participants are given points for each month they are engaged in the study and each month there is a lottery for a large prize or several smaller prizes, with the points acting as lottery tickets and thus the greater the number of points the greater the chance of winning - we will develop this idea further for CD2/3
- Building on the points for participation, we propose extending that system so that participants can earn points for completing tasks, and these points can be exchanged for KHP items or perhaps for other incentives to be determined

(3) We recognize that the respondents will be very eager to see their data, and also that giving participants any type of feedback on their personal information does introduce bias into the study. We will be very mindful of the type and context of the personal information shared with the participants.

- We will provide direct health information in the form of personal anthropomorphic measurements and lab results
- Financial information in the form of year end summaries or topic specific analysis
- We will build into the KHP website some population-level data visualization capabilities

Table 4-1: *Incentive Structure with Possible Examples of HUMAN Project Incentives*

	Monetary value	**Community/trust**	**Feedback/data**
Recruitment	* Set amount for initial visit – same for every household * Set cash amount for each repeated visit to the household before being asked to enroll		
Enrollment	* Set amount for time – same for every household * Smartphone * Smartphone contract * Data plan upgrade for those using personal smartphones	* Selection of KHP items for participants to choose one age-appropriate gift for each household member	* Two-week check in – lab results for confirming that the data collected so far is correct
Retention	* IT support for smartphones * Set small amounts for completion of a task (such as submit copies of tax returns)	* Project holiday & birthday cards and newsletters * Earn points by completing assignments * Interactive portal on the KHP website * KHP merchandise	* Periodic personal data sharing * Resource referrals to explore given personal data such as health or financial planning websites * General study population information results * Interactive data visualization on the KHP website

6.2 Three Stages of Incentive Delivery

There will be three stages of incentive delivery across the lifecycle of the study: recruitment, enrollment, and retention.

(1) Recruitment starts with the selected household agreeing to listen to the introductory information session. The amount paid will be the same for every household and will be paid at the end of the session, preferably in cash. This payment is only contingent upon listening to the session, not for agreeing to participate in the study. Additional incentives will be paid for each subsequent visit before the enrollment team asks the household to join the study.

(2) Enrollment will be the longest and most intense session with the participant household. The household will be given a set cash amount, which will be the same for every study household regardless of the number of residents, and each household member will be able to choose an age-appropriate KHP merchandise as an individual gift.

(3) Retention will be for the life of the study. We will use a mixture of cash, KHP and community good will gifts and information, and personalized and general population data and feedback to keep the participants engaged across the span of the study.

6.3 Incentives at Recruitment

Due to the anticipated challenge of gaining access to the homes of the sampled households, we suggest that a large monetary incentive be given to households that agree to let the recruitment team in the residence and give the fifteen-minute information session. This cash incentive will be given to the household whether or not they agreed to a second visit by the recruitment team or to an enrollment appointment. Repeated visits will be made to households before they are asked to enroll in the study. Cash incentives will be paid for each of these visits, and we may want to reduce the amount of cash paid at subsequent information sessions. We will need to put a limit on the number of incentives paid to households before a final disposition (enrolled/declined) is made, and these details will be developed in CD2.

6.4 Incentives at Enrollment

We will be taking substantial amounts of people's time, which we want to compensate them for. Therefore we anticipate that a large monetary incentive will be paid to the participant household at enrollment in the study. When enrollment and baseline data collection are completed at the same visit, the full enrollment incentive will be given to the household. We are still developing a tiered incentive structure for when the consent and baseline data collection are conducted over more than one visit. When these appointments are spread out, we propose paying a small portion of the full incentive for each completed part of the baseline visit. Conversely, if we give the full incentive once the consent process is completed, we will foster good will and reciprocal behavior.

In order to collect a variety of financial and experiential data, we plan to provide respondents with smartphones loaded with the Kavli HUMAN Project mobile application that is being developed for the HUMAN Project. Smartphones and cellular phone contracts will be given to those participants who do not have a smartphone. We think that the use of a cellular contract as an incentive will also be a deterrent to a participant selling the smartphone. For those participants who choose to use their own smartphone, we propose providing a data plan upgrade to support the use of the Kavli HUMAN Project mobile app. During CD2 we will review all possible alternate incentives for participants who use their own smartphone so that there are not differential amounts or types of incentives being deployed at this stage.

At enrollment, participants will also be given the opportunity to choose one age-appropriate individual gift from the range of KHP branded merchandise.

6.5 Incentives Across the Span of the Kavli HUMAN Project: Engagement and Retention

Cash incentives will be used at the initial stages of recruitment and enrollment, but once a participant is enrolled in the study, other forms of incentives will be the primary source of inducement, such as KHP merchandise, monthly lotteries, and data sharing. Respondents who participate in the 250-member sub-sample will be given a mixture of monetary and other incentives for completing the more in-depth physiological measurements, and these will be determined in CD2.

We recognize that participants will be interested in receiving information about themselves specifically and the HUMAN Project generally, and that this information will be a powerful incentive. We need to be diligent in determining what type of data to disclose and when, so as to not influence the behavior we are tracking. We propose an in-study fragmented feedback experiment with a "year-end summary" of some facet of the personal data to be shared with subsets of participants. For example, a third of the participants may receive a summary of their annual household utility use, another third receive information on their healthcare expenditures, while the remaining third of study participants receive a summary of general household expenditures. To avoid selection bias,

households will be randomly assigned to one of the three treatment groups.

We think that consistent and timely engagement with participants will be crucial to participant retention. We will have a HUMAN Project website, email, and telephone number. Additionally, when a household enrolls, they will be assigned one member of their recruitment team as their specific project liaison, and will be given that person's name and contact information. This project liaison will be their main study contact that will be available to answer their questions and concerns, and will also be the one who maintains contact with the household over the course of the project. For example, if a technology report shows that the household passive data collection dock is no longer sending a signal, the project liaison will be the person who contacts the household to schedule a technology service appointment.

A project newsletter is one way that we plan to keep in touch with participants in a consistent manner. This newsletter could be either electronic and/or a print copy, and the publication frequency could be determined by the amount and type of information to share.

Personal communication will also be an important component of participant engagement. Each year every participant will receive a HUMAN Project birthday card and a holiday card. This communication strategy has been validated by many previous studies as an effective tool for participant retention.

6.6 Anticipated Treatment Effects Associated with Incentives

All incentives introduce treatment effects into the study. Some of these can be adjusted for and some must just be acknowledged as an inevitable source of bias in the conduct of research studies that use incentives.

All the incentives, in the form of money, services, or information, will have treatment effects. Even worse, these treatment effects interact with the different demographics of the subjects. It is paramount to design a set of incentives that provide the minimum monetary compensation and information required to recruit and retain participants. The effects of these incentives on our subjects will need to be explicitly modeled by our econometric staff. Detailed models of these treatment effects will be developed in the CD2 design stage.

Below are examples of the kind of differential treatment effects and challenges we face:

- Providing smartphones to those that do not have them
    - This is a treatment for only some participants who are more likely to be of low SES or seniors
- Some participants, particularly seniors, will have to be incentivized to switch from their current technology
    - We might want to use Bluetooth beacons on a walker or bracelet as a backstop measure for technology resistant subjects
- If we give participants specific personal data (health, financial, genetic) they may change their behavior based on this information
    - The fragmented feedback experiment will give us information to develop methods to correct for this types of treatment effects

7. Attrition and Replacement

We anticipate that the attrition rate will be highest during the first year and then the study population will stabilize, with smaller numbers of study participants leaving the study across time. We will be refreshing the study population every 6 to 12 months to keep it representative of the population of New York City.

7.1 Early Phase Attrition

Knowledge gained from the conduct of other longitudinal studies suggest that during the first six months of enrollment a sizeable number of participants, perhaps up to 25%, will stop engaging in the study. After that initial period, the study population tends to stabilize (Figure 4-6.) During CD2 we will develop a cost-benefit analysis of the level and intensity of effort that would be needed to retain these early phase non-compliant participants.

7.2 Missing Data Points

The optimal experience will be for every study subject to participate at each data collection point. However, for a variety of reasons, we know that there will be times when the data requested will not be collected. This is an understandable and anticipated issue for the study. We will be monitoring data collection carefully and will attempt to minimize missing data through close contact with the study participants.

7.2.1 Non-Contact with Project Participants

There will be a small percentage of enrollees who will stop actively participating during the course of the study. We will continue the passive data collection on these individuals, yet there is a cost benefit analysis to be done to determine the threshold at which we will stop pursuing participants who no longer respond to data collection requests. Based on the data from the U.S. HRS, it is estimated that a large fraction of the budget could be spent to retain a relatively small fraction of the participants (~10%), so decisions must be made about whether there are resource limits in preventing this kind of attrition.

7.3 Participant Withdrawal from the Project

It is anticipated that over the course of the study a small number of participants will request to be fully withdrawn from the Project. As the study relies

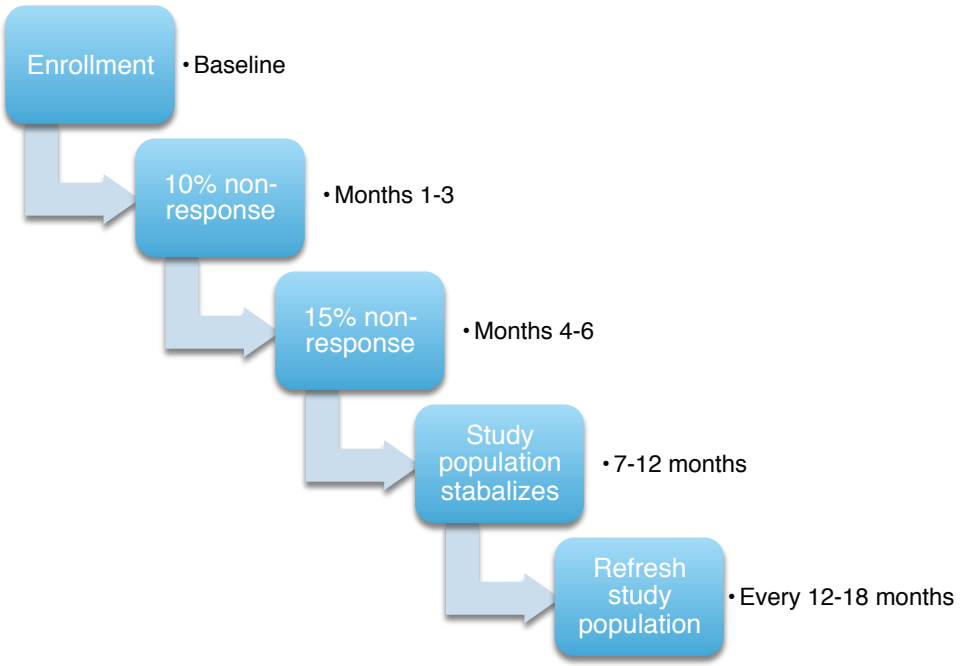


Figure 4-6: *Estimated Study Population Attrition and Replacement*

extensively on automatic data collection, it may be possible to avoid having people entirely disappear from our data logs, even if they choose not to actively participate in the study after some point. To make this possible, at initial contact it will be essential to obtain consent to continue passive data collection about our subjects even if they choose to withdraw. Ideally, we will be able to track subjects even after they stop actively responding through Social Security records, Department of Education records, Statewide Planning and Research Cooperative System (SPARCS) data, IRS data, and Veterans Administration data, for example. In this way, we should be able to continue to collect data about our participants, even if they stop participating in the study. Of course, participants have the option to entirely opt out of further data collection, but it is assumed that this form of absolute withdrawal will be rare.

7.4 Out-migration

A common life change for which we will need to be prepared is a move out of New York City. While demographic data suggests that a representative cross-section of the city will yield subjects who overall move very little, higher income groups, in particular, do move. For the study's purposes, moves can be divided into four categories, each of which we address with a different strategy:

- *Suburban Out-migrants*: People still in the New York metropolitan area – we will keep these participants in the study as they probably still come to the city occasionally and so should not be too difficult to schedule for a physical follow up session. These subjects will not be replaced.
- *Temporary Out-migrants*: People temporarily out of the NYC area such as snowbirds and college students – we will keep these participants in the study and schedule follow up sessions when they are in the NYC area. These subjects will not be replaced.
- *National Out-migrants*: People who move out of NYC metropolitan area but are still in country – we will collect the shed data such as government records, and continue to push surveys and collect cell phone data. These subjects will be replaced.
- *International Out-migrants*: People who move out of the country. No data collection will be attempted for these participants. These subjects will be replaced.

7.5 Transitional Life Events

As younger members of the residential groups turn 18 years old they will need to complete the HUMAN Project informed consent for themselves. This will be a critical transition as these subjects may be particularly sensitive to withdrawing from the study at this time. We will need to develop specific recruitment/retention materials for these young adults.

People will also change life circumstances, going to places where it will be difficult to participate in data collection such as college, nursing homes, become homeless, or enter the prison system. Some of these demographic transitions may be easier to plan for; for example it may be possible to schedule appointments with college students when they come home for school breaks.

A percentage of our total study population will enter the criminal justice system each year. It will be of critical importance to follow these participants as much as possible as they move through the legal system and correctional facilities. We will work to determine the steps along the way where we can interact with the participant to capture personal data, and to setup mechanisms to gather institutional data. As the participant transitions back into the community, this would be an opportunity to resume regular data collection.

Additionally, some individuals and families will become homeless during the course of the study. This may mean moving in with family or friends or formally entering the New York City shelter system. We will be in contact with participants though their smartphones, but will no longer have any data captured by the household docking station. If a participant enters the shelter system, we will be able

to get metrics such as the dates of entry and exit from the system, and the number of nights in the sheltered.

These are among many circumstances we foresee where it will be necessary to do a cost-benefit analysis in CD2. We need to determine the level of project resources that should be devoted to tracking people under difficult conditions, balancing our desire for information with respect for the respondents' privacy.

7.6 Subject Replacement

The study population will be dynamic over time to reflect the changing demographics of the city. As people die or leave the study, new participants will be selected to replace those who are lost based on the sample design of the Master Study Frame. The HUMAN Project demographer will be responsible for model maintenance. The sample demographic updating processes will be automated and the sample frame will be refreshed every 6-12 months. In this way, the study population will continue to reflect the ever-changing population of New York City as much as possible.

8. Project Administration

8.1 Protection of Human Subjects in Research

We anticipate that the New York University Faculty of Arts and Sciences Institutional Review Board will be the IRB of record for the entire HUMAN Project. HUMAN Project staff will be responsible for the initial IRB application, all revisions, annual renewal, and for keeping the project in compliance with all NYU regulations.

We recommend that all study personnel who will have contact with participants or participant data, complete an online protection of human subjects in a research training course, such as the one offered by The Collaborative Institutional Training Initiative (CITI Program) at the University of Miami https://www.citiprogram.org. We will develop an unanticipated problem and adverse event protocol following the guidance laid out by the U.S. Department of Health and Human Services.

Although procedures during the intake procedure would help to identify mental health problems in participants at initial screening stage, it is also expected that additional mental health issues will arise over the course of the study; the frequency of such diagnoses could be as high as 1 in 5 young people. These subjects may require specific assets from the study, such as a referral to a mental health professional or modified data collection schedule if hospitalized after the onset of mental illness. We also expect that a subset of the seniors in the study will develop dementia. Procedures will be developed for handling the consent (and possibly re-consent with a proxy) process in these situations, as well as for facing attendant difficulties with data collection in dementia patients.

Referrals for mental health and social services, such as toll free hot lines and city-specific agencies, will be given to all respondents who request assistance. We are considering having a licensed mental health professional on-call for emergency situations that we may encounter during the study.

8.2 Legal Considerations

There is the possibility that in the normal course of conducting the HUMAN Project we may uncover suspected child or elder abuse and neglect. We take these issues very seriously and will work with the legal department of NYU to develop a protocol for project staff that detail requirements and the specific chain of command for reporting suspected abuse. We also propose to develop study procedures that will flag participants for referral to a study team of physicians, neurologists, and psychologists. Finally, we are considering recruiting a HUMAN Project study ethicist.

We will apply for a Certificate of Confidentiality (CoC) from the National Institutes of Health. CoCs are issued to institutions or universities where the research is conducted and are meant to protect the

privacy of research subjects by protecting investigators and institutions from being compelled to release information that could be used to identify subjects within a research project. Researchers can use a Certificate to avoid compelled "involuntary disclosure" and to refuse to disclose identifying information in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level, the names and other identifying information about any individual who participates as a research subject any time the Certificate is in effect. It does not protect against voluntary disclosures by the researcher, but those disclosures must be specified in the informed consent form.

Identifying information protected by a Certificate may be disclosed under the following circumstances:

- Voluntary disclosure of information by study participants themselves or any disclosure that the study participant has consented to in writing, such as to insurers, employers, or other third parties;
- Voluntary disclosure by the researcher of information on such things as child abuse, reportable communicable diseases, possible threat to self or others, or other voluntary disclosures provided that such disclosures are spelled out in the informed consent form;
- Voluntary compliance by the researcher with reporting requirements of state laws, such as knowledge of communicable disease, provided such intention to report is specified in the informed consent form; or
- Release of information by researchers to DHHS as required for program evaluation or audits of research records or to the FDA as required under the federal Food, Drug, and Cosmetic Act (21 U.S.C. 301 et seq.)

9. Kavli HUMAN Project Pilot Test

We propose a pilot test towards the end of the final (CD2/CD3) design stage. This will be to ensure that our recruitment plans will provide a sufficient yield for the study population, that our enrollment steps function as intended, and to run through the bio-specimen collection process.

10. Draft Data Collection Timeline

Assumptions:

1. Project will be conducted in English and Spanish only
2. There will be 3 HUMAN Project vans – two teams of recruiters and enrollers per van
3. Sufficient and consistent field staff
4. Data collection 50 weeks per year

Table 4-2: *Estimated Timeline of Project Launch*

Months

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Sampling method developed																		
Construction of sample frame																		
Development of project materials																		
Translation of project materials																		
IRB submission																		
Hiring project field staff																		
Training project field staff																		
Purchasing equipment																		
	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Hiring project field staff																		
Training project field staff																		
Purchasing equipment																		
Sample updated and drawn																		
Pilot test																		
Refine materials/process																		
Project Launch																		
Mail post cards and packages																		
Deploy to neighborhoods																		
Data collection																		

Table 4-3: *Data Collection Estimated Numbers*

Recruitment level of difficulty	% Sampled households eligible – based on language ability	% Sampled and eligible households that agree to participate – based on response rates for NYC HANES	#Households enrolled per week/per van – the total in parentheses is based on 3 vans in the field	#Households enrolled per year – based on 50 weeks of data collection per year and 3 vans in the field	Approximate number of weeks in the field	Total months of data collection
Easy	**90**	**30**	**10 (30)**	**1,500**	**84**	**21**
Moderate	**88**	**25**	**7 (21)**	**1,050**	**120**	**30**
Hard	**85**	**20**	**5 (15)**	**750**	**167**	**42**

The data collection table (Table 4-3) is an estimate of the challenges to recruitment based on language fluency, response rates of selected households, the number of project vans with complete recruitment and enrollment teams in the field, and a data collection period of 50 weeks per year. These estimates were made using information gathered from NYC HANES, NYC HVS and other longitudinal studies. The recruitment level of difficultly are to be read horizontally left to right, and in actual operation it is anticipated that there will be a mixture of variables: such as if we use instruments written in "Simplified English" this may raise the percent of language eligible households to higher than 90%, and we anticipate cooperation and response rates will fluctuate between boroughs and neighborhoods, and even times of the year, and some weeks a van may be out of service.

NYC HANES was a population-based, cross-sectional survey of NYC adults using three-stage cluster sampling. Between August 2013 and June 2014, selected participants completed a survey and physical exam. Of the 3065 households approached, 2742 were eligible and 1827 were successfully screened (67%). A total of 1524 of eligible participants completed the survey (54%), for an overall response rate of 36%[1]. We think that the HUMAN Project will have a slightly lower overall response rate than NYC HANES due to the more in-depth and longitudinal nature of the study. However, the HUMAN Project will be using a very similar sampling strategy and we think that this study represents a reliable starting point for estimating response rates.

We reviewed the New York City Housing and Vacancy Survey (NYCHVS) because it also uses an address based sampling method that is similar to the one proposed for the HUMAN Project. NYCHVS is sponsored by the New York City Department of Housing Preservation and Development, and is

[1] Thorpe LE , Greene C, Freeman A, Snell E, Rodriguez-Lopez JS, Frankel M, Punsalang A, Chernov C, Lurie E, Friedman M, Koppaka R, Perlman SE. Rationale, design and respondent characteristics of the 2013–2014 New York City Health and Nutrition Examination Survey (NYC HANES 2013–2014). *Preventive Medicine Reports*. 2015;(2):580-585.

conducted every 3 years to comply with New York State and New York City's rent regulation laws. Sample units for the 2014 NYCHVS came from two primary sources: 1) the 2010 Decennial Census files, and 2) a file of addresses listing all residential units, citywide, issued Certificates of Occupancy for new construction from April 1, 2010 through November 30, 2013. Approximately 19,000 units throughout the city were selected as a representative sample of the housing in the five boroughs of New York City. Each sample unit represents approximately 170 similar housing units. The Census Bureau attempts to obtain an interview at each sample unit. The interview rate for the 2014 NYCHVS was 92%[2]. We believe that the very high response rate for the HVS is due to fact that it is mandated by the City of New York and conducted by the US Census Bureau, which may give the impression that residents are required by law to participate.

11. Project Directions to be Further Considered

1. Residential unit member participation

(a) Do we require that all members of a household agree to participate in the study for any members to participate in the study?

(b) Do we enroll a household if one of the members is ineligible?

(c) Should the minimum amount of time a person must spend as a resident of the household in order to be eligible as the "seed" participant be 50%?

(d) Should the minimum amount of time a person must reside in the household to be counted in the residential unit be 5%? This amounts to about 18 days a year, should the percentage of time spent in the household be set higher?

(e) Does the seed have to agree to provide information on non-residential family members?

(f) When an established residential unit separates, how do we determine which members to follow?

- If the seed remains at the sampled household, do we limit data collection to the seed and the residential unit members remaining?
- If the seed leaves the sampled household, do we continue data collection for just the seed? What about other residential unit members the seed may take with them?
- What if the seed is a minor child of divorced parents with shared custody?
- Need to re-consent children as adults when they turn 18 years old.
- Can these former child participants remove their old data if they choose not to re-consent?

(g) Should children receive an incentive? Should each participant in the residential unit receive the same incentive, and the parent/guardian receive the incentive on behalf of the children younger than 18?

2. Non-residential family member participation

(a) We need to develop this idea to determine what we want to collect from these extended family members. Mailing self-administered saliva sample kits with a short questionnaire could be a first step.

(b) Must this information be gathered at baseline? We may want to give the seed and residential network time to experience the study before requesting referral information.

[2] http://www.census.gov/housing/nychvs/

3. Languages to be used for the project

(a) Do we conduct the project in English and Spanish only? This will capture approximately 88% of households.

(b) What criteria will we use to determine that non-native English speakers understand the requirements of the study and can give informed consent?

(c) Suggestion to conduct the study in: Standard English, Simple English (6th grade level), and Standard Spanish.

(d) Should we include Mandarin, which will increase coverage of the city to approximately 92%, but will require unique staff and training resources that will add much more time to development and testing and may not yield that many enrollees?

(e) Should we develop plans for the inclusion of a third (Mandarin) or fourth (Russian) language for future waves of data collection?

4. Recruitment and enrollment

(a) We planned that our recruitment staff leadership will attend the recruitment training conducted by the Institute for Social Research at the University of Michigan, and that all data collection will be conducted by permanent or temporary HUMAN Project staff. Is this the strategy we want to pursue or do we want to consider sub-contracting any portion of data collection?

(b) How can we accommodate people who do not want us to conduct the in-take assessments in their house? Is there a lab or exam rooms we can use as needed in any NYU campus?

(c) When discussing HRS with Nicole Kirgis, she stated that in the 2010 HRS they used a two-step recruitment process and had lower than expected response rates and many scheduling issues. We may want to consider having the recruiters trained to administer the informed consent, the household census and basic demographics, and this could be set as the minimum amount of data to be collected to be considered a completed case.

5. Triggered data collection

(a) From the Advisory Council meeting there was some disagreement about the use of triggered data collection when a respondent reports the occurrence of a major life event at a regularly scheduled assessment.

- An advantage would be to increase the amount of information gathered at a particularly interesting point in the respondent's life, such as when a baby is born or a person becomes unemployed.
- Disadvantages include how the non-uniform data collection could impact the statistical model, and issues related to what might be appropriate triggers, and would this type of data collection be intrusive during a sensitive time in a respondent's life.

6. Crowd-source population

(a) This could possibly be a very large source of data from as many as 300,000 New York City residents. Below are a few of the many issues to consider:

- We could engage many interested citizen-scientists in the study and this could increase the power of the measurements.
- This would be a sample of opportunity and the selection bias could substantially limit the generalizability of the observations.
- This could possibly be great for public relations and public engagement with the research if it is managed very well.
- Keeping this data safe could be difficult and expensive.

7. Data discrepancies

(a) We need to develop policies to identify and resolve conflicts between self-report and administrative data.

(b) We need to develop policies to identify intentional fraud or misrepresentation and how we would remove these participants from the study.

8. Observational data to be included

(a) During the course of the study there will be many policy changes and one way to increase the inferential power would be to look at the effects of policy changes, particularly those limited in either time or location.

- We need to determine which types of observational data we want to track and how to capture it in our database.

9. Legal issues

(a) We need to work with the NYU legal department to develop protocols for suspected child or elder abuse or neglect that are aligned with federal reporting requirements.

(b) Can we have a staff of on-call medical professionals and flag for referral respondents who exhibit behaviors such as:

- Express suicidal ideation on a depression scale
- Have critical lab values
- Appear to be losing cognitive function?

(c) Other than intentional fraud, what are other actions that we think should result in a participant being removed from the study population?

(d) Should we recruit a study ethicist?

10. Prison and homeless populations

(a) We want to continue to follow people who enter the jail/prison system and those who become homeless. This will be a challenge to continue data collection in a way that is systematic yet sensitive to the personal situations of the participants. We will need to develop protocols for maintaining contact with participants in these circumstances.

11. Field staff retention and duties

(a) We anticipate that once we reach the goal of 2,500 "seed" participants and enrollment slows down to subject replacement, we will need to let go of recruitment specialists. These individuals will likely be the project contact liaison for a number of households. How should we handle this situation?

(b) At the height of initial project enrollment, there will be 6 enrollers. We could retain the 3 of the best enrollers for the replacement enrollment, and transfer the project liaison duties to these individuals.

(c) Other support staff will be assigned as needed to the project liaison role.

12. Incentives and incentive structure

(a) We need to develop a suite of options for monetary payments, including cash, electronic transfer (PayPal, Western Union, bank account), gift cards, and other methods.

(b) We need to set up the incentive structure for the in-depth physical measurements that will be done with the sub-sample of 250 participants.

5

PRIVACY AND SECURITY

1. Introduction

The Kavli HUMAN Project (KHP) will create a resource for scholars that can be used to address a wide range of basic research and policy questions. However, this comprehensive data set will contain a wide range of sensitive material, including personally identifiable information, so access must be balanced against the need to protect the security of the data and the privacy of the participants. New technologies make it ever easier to ensure appropriate data security controls, and a reliance on strong data governance policies and bodies to implement them will support privacy protections. Successful management of the rich store of electronic data and biological samples is critical to the success of the project, and thus we have invested substantial resources in the design and implementation of this aspect of the study.

2. A Framework for Designing the HUMAN Project Privacy and Security Controls

General guidance for data management and governance policies is under the auspices of the Privacy and Security Advisory Council, and ultimately, the Project will require a dedicated Chief Privacy and Security Officer. However, given the specialized technical and regulatory knowledge necessary for designing appropriate data security controls for the KHP data facility, even at this early stage, a report was commissioned from Hogan Lovells US LLP, a law firm with a world-renowned Privacy and Information Management practice. The report was written by: Marcy Wilder, director of this group and a former Deputy General Counsel of the U.S. Department of Health and Human Services, where she was the lead attorney in the development of HIPAA privacy and security regulations, as well as co-authors Paul Otto and Brian Kennedy from Hogan Lovells. The key points in the report are highlighted in the following pages, and the full report can be found at the end of this document as Appendix F.

The data collected under the auspices of the KHP will be subject to a number of legal and regulatory requirements. A primary source of oversight for the KHP will be the New York University (NYU) Institutional Review Board (IRB), which will approve the study design prior to implementation. Among the priorities of the IRB are minimizing participants' risks, ensuring their safety, protecting their privacy and ensuring that the confidentiality of the data is maintained. New York State does not currently have any specific legal requirements for protecting personal information, though there are data disposal laws that will govern the destruction of any records of personal information. Although the KHP is not technically a "covered entity" under HIPAA regulations, nor legally required to adhere to the Payment Card Industry Data Security Standard (PCI DSS) or the Gramm-Leach-Bliley Act that regulates the financial industry, these rules are becoming the industry standards for managing health and financial information, so it would still be appropriate to meet these requirements. In contrast, New York State student data privacy laws govern any third parties that receive student data, so all of

the KHP education data would need to conform to these regulations.

Hogan Lovells has suggested that a way to address this patchwork of regulations would be to take guidance from the National Institute of Standards and Technologies (NIST). According to their report: "NIST Special Publication 800-53, *Security and Privacy Controls for Federal Information Systems and Organizations* provides a comprehensive catalog of security and privacy controls and outlines a process for selecting appropriate controls based on the mission, size and threat profile of an organization." This security framework (available in full at http://csrc.nist.gov/publications/drafts/800-53-rev4/sp800-53-rev4-ipd.pdf) was designed to apply to a wide range of information systems and has incorporated security controls from all sectors of government activity and to facilitate compliance with all applicable federal laws, regulations and guidelines. However, it is noteworthy that there are additional requirements for HIPAA and PCI DSS compliance that are not incorporated into the NIST framework, but, as discussed above, will still need to be considered as part of the design of security control policies.

The security controls in NIST SP 800-53 are organized into three basic areas: management controls, operational controls and technical controls. Management controls include high level controls across the entire KHP, such as program management, risk assessment, planning, system and services acquisition and security assessment and authorization. Operational controls include personnel security, physical and environmental protection, contingency planning, configuration management, maintenance, system and information integrity, media protection and incident response, as well as awareness and training. Technical controls cover identification and authentication, access control, audit and accountability and system and communications protection. However, note that although the NIST guidelines provide a comprehensive framework for addressing privacy and security controls, they do not specify the methods required to achieve these controls. This will allow us to choose an optimal solution specific to the KHP. With the additional inclusion of controls to address HIPAA and PCI DSS requirements, addressing each of these areas will lead to the development of a comprehensive plan for protecting the privacy and security of the Kavli HUMAN Project participants and their data.

The Hogan Lovells report also makes strong recommendations about the process for developing a comprehensive information security program with a focus on integrating risk management processes across the board. They recommend a six-step process for moving from defining leadership responsibilities, to identification and inventory of data and systems, to security categorization. Once these aspects have been specified, one can select appropriate security controls, define security policies and procedures, and finally, implement, monitor and test the systems. We will use this process to complete the next phase of the study design in which we begin to specify these key details.

3. A Preliminary Design Plan to Meet Privacy and Security Control Requirements

Although we have not yet reached the stage where we can specify detailed strategies, based on the Hogan Lovells report, we describe the preliminary design for a data facility that would meet these requirements. This preliminary design lays out a basic plan along the lines suggested by Hogan Lovells, which we will continue to refine as the project takes shape and more detailed planning can be implemented.

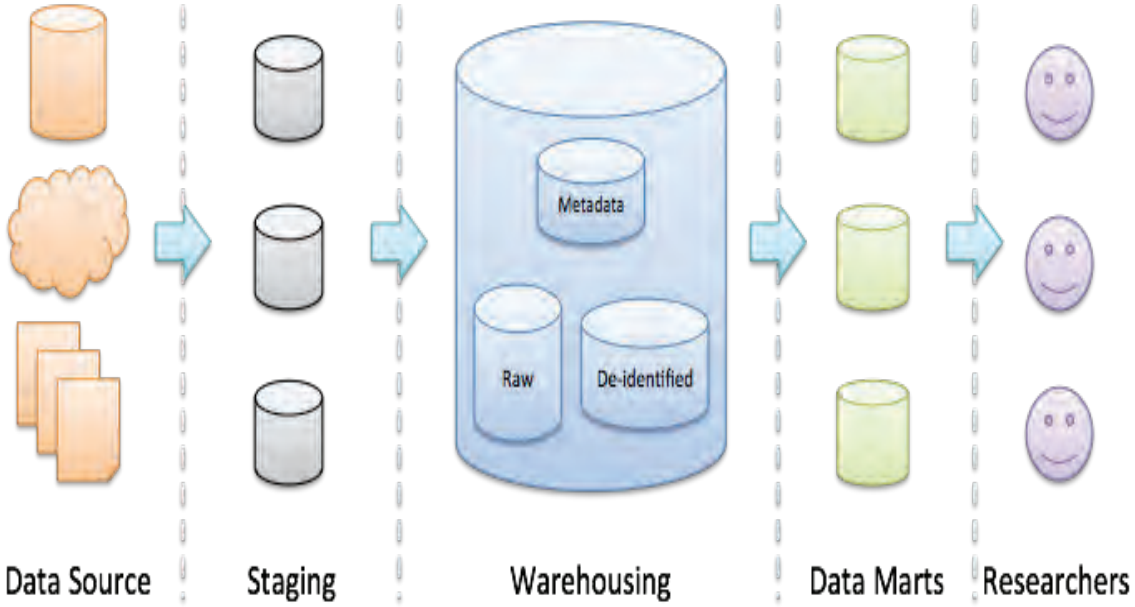


Figure 5-1: *CUSP Data Facility Architecture*

We propose to implement a data platform similar to the data facility that CUSP is currently building (Figure 5-1). Data goes through a staging process as it is ingested into the warehouse for storage. Then, when researchers want to run queries or perform analyses, specialized temporary "data marts" are created which contain only the relevant data sets. As indicated by the arrows in Figure 5-1, data can only move in one direction through the data warehouse, so that unauthorized access to the warehouse cannot be achieved from the staging server or the data marts. This will be achieved through both electronic means as well as physical means (simply cutting the pins on the connection cables and creating a "data fountain").

Three features of the data marts enhance the security of the data. Each data mart is created for a specific project, and researchers have access only to the data mart(s) required for their own project for the amount of time required to perform the necessary analyses. After the researcher is finished, the data mart is deleted (though the data itself always remains safely stored in the warehouse) so that it is not vulnerable to unauthorized access. The data marts themselves are thus heavily partitioned silos.

Some additional security measures will also be necessary because data will be coming directly from our participants' smartphones or other devices outside our direct control. In order to avoid contamination, it will be necessary to scan for malware and viruses prior to ingestion. The applications on the phones themselves will need to be written with a way of checking the integrity of the incoming data.

Network segmentation is also a critical feature of the system design. Security breaches are unavoidable despite even the most stringent approaches, and splitting up the network can help improve security. For example, it may prevent an intruder from gaining broad access with stolen network credentials. Creating sub-networks or layers can also make the network more robust by limiting the effects of local failures on other parts of the network.

3.1 Access Control

In addition to the protective structure of the data warehouse architecture, additional physical and electronic access controls will be necessary to ensure the security of the HUMAN Project database. Given the value of the data and its highly sensitive content,

it will be a very attractive target to malicious attackers, and "hacktivists" might attempt to disrupt the database to make a point, so it is critical that strong protections are put in place. Below are some of the approaches that we are considering implementing as part of our security plan.

We will, of course, implement the standard access controls for database security. Basic physical access controls used for sensitive data stores include securing the server rooms and limiting the number of people allowed to physically access the servers. Requiring double-identification methods for physical access to data storage systems (i.e. finger print and access card) is also likely. Physical access limits could also be imposed by air gap – a double firewall created by requiring researchers to use kiosks that access only the data marts, though this would likely only be practical when users wish to access the most sensitive data sets.

Electronic access should generally be granted on an as-needed basis, and broader access limited to a relatively small number of people who have passed strict background checks and regulated by role-based access privileges. Double-identification systems will also likely need to be implemented for electronic access (i.e. fingerprint and password). For mission-critical functions, like those that could imperil the security of the entire database or the large-scale privacy of our subjects, "two-key" systems will likely be required. In such a system, only when two system operators, who have both consented and been double-identified, approve of an action, can a function of this category be implemented. (Sony's recent hacking by North Korea rested on its failure to use any two-key systems at all.)

However, even with careful administration of access privileges, it will be important to implement additional steps to protect the data from malicious insiders (or someone stealing the credentials of an insider's account, for lower security single-identification functions). Measures such as login monitoring and behavioral monitoring of administrators using automated systems that look for 'out of the ordinary' behavior can be very powerful and are standard in most high security data environments. Such systems will need to be implemented as well. It should be noted that these automated alerts for unusual patterns of behavior work best for long-term employees who can be easily and closely monitored, and for whom there are clear expectations of behavior. If we expect to allow scholars to access the database directly, it may be more difficult to use these approaches. For this reason (and others), it is unlikely that scholars will be allowed direct access to the actual database. Only when the highest possible level of security is operating, can critical privacy and security functions be accessible if we hope to maintain the degree of security necessary for this project.

Another potential source of protection might be a "gateway solution," where bandwidth control is used to restrict how much data can be moved off the server at one time. However, some experts question the strength of this strategy, as recent security breaches have been executed by moving data in packets just a bit below the transfer threshold. Still, as an additional measure, this could prove valuable.

3.2 Access and Use

Even the most perfectly secure data system, however, is only as secure as the uses to which it is put. The use of secure data marts and a state-of-the-art security environment can largely guarantee that only permitted uses of the data occur. But in order to assure the privacy and security of our subjects, it is essential that the study management govern the use of data effectively. To achieve that, there will need to be a committee and a process to evaluate who is granted access to specific classes of data and under what terms. This will be particularly important for defining the access privileges granted to visiting scholars and others who may initially be unknown to the study team – and thus will have limited physical and electronic access to our systems. Important criteria to consider include (but are not limited to) security clearance procedures for scholars and employees, credentialing and background checks. Institutional review boards (IRBs) at researchers' home institutions could, in principle,

provide some level of assistance with the initial vetting process, as it would not be possible to use the database for research without an IRB approved protocol. However, IRBs do not have sufficiently uniform standards to provide adequate information for making access decisions, so additional evaluation by a committee under the aegis of the study will still be necessary. A basic outline of the Credentialing Committee will be required for the CD2 document and a detailed overview of its policies and procedures will be required no later than the CD3 document.

Even after substantial vetting, external researchers, particularly those working off site, are likely to be the greatest vulnerability to the system. Training in security and data management prior to any data access will be critical to ensure that researchers comply with all policies and regulations. Once work with the data commences, activity of the researchers on the network will need to be closely tracked and access privileges re-evaluated every quarter.

An additional approach would be to encourage collaborations by external researchers with existing staff researchers. Such collaborations would provide additional protection by limiting the number of people who can access the data, but could also add value for researchers (particularly those who are inexperienced with handling large, sensitive data sets) by offering an opportunity to learn from an expert. This approach, while likely improving the quality of research using the database, requires a sufficient quantity of expert resources in-house to support a reasonable number of collaborative projects. We note that this model is similar to the work of "trusted telescope operators" in national astronomical observatories – specially trained engineers that facilitate the data gathering operations of large-scale telescope facilities. However, it is important to note that this approach requires the financial resources to support technical staff. (We note that this could be funded via data use fees of some kind, perhaps employing a sliding scale by which researchers with limited funds are subsidized by for-profit research entities.)

3.3 Special Protections for Especially Sensitive Data

As is the standard protocol at CUSP and many other places, all data ingested into the database will be classified along the continuum from highly sensitive (and thus most stringently access restricted) to not at all sensitive (and thus publicly available). This provides the opportunity to increase security measures for the most sensitive data without creating excessive barriers to the use of less sensitive data. This may mean adding additional layers for separating sensitive data, triple de-identifying it with different codes for different types of data and using a 2-key system to access the most private sets of data. Separating out sensitive data with different levels of security would be necessary at the level of both data storage and data marts.

3.4 Disaster Survivability-recoverability

Two general strategies are commonly used to promote data recovery following disaster: mirrored servers and back up to a separate physical medium that is subsequently stored offsite. We note that standard procedures break the datasets into a large number of encrypted components, each typically stored at different locations, which must be brought back together for reconstruction of the database. Such an approach is essential for securing off-site backups of any kind.

One traditional approach that employs this method has been to create a set of fragmentary data tapes, for example a set of 10, that are then stored in safe deposit boxes at multiple offsite locations. The encrypted fragmentary tapes are each useless alone. A minimum subset, for example any 7 of the 10, are required for database reconstruction with this method. One advantage of such an approach is that physical tapes are less accessible to hackers, since the stored data is not accessible online. However, that lack of accessibility also limits the availability of the back-up media if a restoration is required (particularly if tapes are stored in geographically distant locations). Physical tapes are also a weakness, as experts reported that despite the best

intentions, they have a tendency to get lost. (This compromises recovery more than security for a properly encrypted system that requires that 7 tapes be brought together from different locations for reconstruction.) Costs for this strategy include the purchase, storage and destruction of the physical tapes.

Back-up via point-to-point transfer to remote servers requires that data be sent from the main server to a remote site with very high-level encryption and 2-factor authentication, often with a 2-key system required for recovery. Under this model, data is physically encrypted inside our secure facility and only leaves after that encryption process is complete. Keys for decryption are handled as physically secured objects requiring 2-key double authentication. This method has complementary strengths and weaknesses to physical backups. Because information is being transferred and stored online, it is vulnerable to hackers should they be able to defeat all of the security systems that are in place. However, the backup copy (or copies) cannot be physically lost, and it is more easily accessed if any restoration is required. In the case of the HUMAN Project, there may be some synergies with infrastructure currently being developed by CUSP and Internet2 that may argue for the remote encrypted server approach. It should also be noted that the Privacy and Security Advisory Council expressed some consensus that this remote online server approach will likely be preferred. An initial plan is presented in this CD1 document. For either solution, it will be critical to complete detailed capacity planning to ensure that the chosen option is scalable to our needs in the CD2 document.

In either case, we will also require the development of a process for the creation of back-ups and restoration in case of disaster. The more complete (but consequently more expensive) choice is to have a hot site – this is a location where operations can be resumed after a disaster as soon as you can get people into seats at the new site. A less expensive option is a cold site, where the company that maintains the offsite back up is responsible only for restoring the data itself – the owner is responsible for providing a physical site where operations can resume. A cost-benefit analysis of all options will be necessary at the CD2 stage in order to determine the optimal strategy to ensure protection of the data from disaster and facilitate restoration of the data as smoothly and rapidly as possible.

3.5 System Testing

Once the security controls have been established, regular monitoring and testing of the system are both good practice and a legal requirement. Program managers must assess any changes to project design, regulations or technical vulnerabilities that may arise during the course of the project. Penetration tests should include attempts at both physical and electronic attacks by highly skilled testers.

There will be a core group of security staff members to manage cybersecurity protocols for the project, but program managers may choose to rely on external expertise for some services, such as penetration testing. We have discussed the possibility of employing a company such as Synack, which is essentially a "red team" for hire. They have a staff of skilled hackers who provide ongoing vulnerability testing, and immediately convert any identified weaknesses into specific recommendations for improved security. Such a service can cost about $300,000USD a year (see business proposal from Synack, Appendix G), but given the critical importance of ongoing vulnerability testing could be a worthwhile investment.

4. Privacy

Discussions have largely been focused on privacy and security of the electronic database that will be associated with the study. However, it will be important to also consider the privacy and security of all stored biological samples. Currently we have proposed to store samples of biological material such as cell lines, genetic material, blood, urine and hair. Although the protection of physical samples was not discussed in detail at the Privacy and Security Advisory Council meeting, a separate plan

will be necessary for biological samples and it must meet equally stringent requirements for privacy, security and disaster protection as the plan that will be designed for protecting electronic data. Physical samples will likely be stored in a commercial facility. The security standards for physical sample storage should be developed for the CD2 document.

An issue that has been raised in several different contexts is the separability of risks from privacy/security breach and public relations/trust problems. These are related – certainly the former will likely result in the latter. However, there are circumstances in which policies that meet legal requirements might still result in outcomes that weaken the trust of participants and/or the general public in the project. Therefore, it is critical to consider the consequences of all potential policies in light of both views.

Data access rights are still in flux, but the legislative environment is evolving quickly, and given the proposed strategy to provide only partial feedback as a way of controlling for treatment effects, it will be important to be transparent about the limitations we will place on participants' ability to access their own data. A new Consumer Privacy Bill of Rights and a new Student Digital Privacy Act (based on California legislation – Student Online Personal Information Protection Act) are expected to be unveiled relatively soon.

A number of data types are protected by specific government regulations: data from health care providers by HIPAA (Health Insurance Portability and Accountability Act), data from K-12 educational institutions by FERPA (Family Educational Rights and Privacy Act) and data from financial institutions by GLBA (Gramm-Leach-Bliley Act). Since the Project is not run by any such entities, there is no requirement to be in compliance with these regulations (which are quite onerous). It has been suggested that it would make more sense to be "in accordance" with such regulations, which would mean meeting those standards, but not bearing quite as heavy a burden. For example, this is how insurance companies function, allowing them to give representation and assurances to people that there is no additional risk beyond what they already experience.

It will be important to look out for the interests of the participants in the study as decisions are made about the use of study data as well as changes to the study over time. One important way to address this would be to add an advisory council made up of participants – a possibility also raised by the Subject Pool Advisory Council. Members of this committee would be chosen to reflect the diversity of the subject pool, and the chairs of that council would serve to represent the participant perspective to other decision making groups for the study.

A more detailed discussion of consent follows below, but one consent-related privacy issue is mentioned here. Among the plans for expanding the study, it has been proposed that investigators might initiate independently funded projects that capitalize on core participant groups. Additional consent processes would likely be necessary for such initiatives, but it should be noted that the study staff would have to manage these additional consents in order to protect participant identity.

4.1 Third Party Requests

Any valuable database will be the target of third party requests, and the HUMAN Project data could be attractive to a number of entities, including, but not limited to, government officials, police officers and divorce lawyers. These requests could be frequent and extensive, though it is difficult to gauge at this point. However, it will likely be necessary to allot budget for legal fees and staff for handling requests.

Several strategies for reducing vulnerability to (and the costs of) third party requests have been proposed, but dismissed as providing insufficient protection. If the database were to be under the auspices of the federal government, it would be protected from civil litigation, but would still be accessible by FOIA requests. Moving the data offshore is also not a solution, because the standard of the law is that if you can access the data for

scientific purposes, then you are obligated to access it upon legal request.

It is therefore best to limit stored data to what is absolutely necessary. For example, it would be ideal to remove identity data entirely from the database; this would provide protection from both subpoena and hackers. Aggregation would have a similar effect – once the data is mixed, data about any individual could not be extracted. However, neither of these are straight forward solutions for this particular project – it is difficult to envision a mechanism that could be used to continually accrue data without an identifier linked to real-life identity at some level. One suggestion would be to have a separate team to do all participant interaction and to keep this team away from the data so that there is a plausible separation. In this way, either team alone would not have the capacity to match data with identity, and the groups might not be legally compelled to cooperate if they were separate entities (although this latter condition was not certain). If a participant fingerprint was required for linking certain kinds of data (e.g. re-identification), this might also strengthen the ability to withstand legal challenge, since the participant would be the only one able to provide the requested data.

However, a de-identification solution would still leave the data vulnerable to requests where the requestor already knew something about the records being requested (for example: "please provide all data pertaining to the individual who was at location X at time Y on date Z"). Separating the control of individual data stores for identity, location, genetics and other sensitive data would provide the opportunity to put in place additional protections for such data sets. Such separation would certainly be useful for strengthening security measures, but is unlikely to be useful in blocking legal requests for data.

Given all of these constraints, the optimal solution for this project is to obtain a **Certificate of Confidentiality for Health Research (CoC)**. From the CoC guidelines:

> Certificates of Confidentiality are issued by the National Institutes of Health (NIH) and other HHS agencies to protect identifiable research information from forced or compelled disclosure. They allow the investigator and others who have access to research records to refuse to disclose identifying information on research participants in civil, criminal, administrative, legislative, or other proceedings, whether federal, state, or local. Certificates of Confidentiality may be granted for studies collecting information that, if disclosed, could have adverse consequences for subjects, such as damage to their financial standing, employability, insurability, or reputation…
>
> Certificates of Confidentiality protect subjects from compelled disclosure of identifying information but do not prevent the voluntary disclosure of identifying characteristics of research subjects. Researchers, therefore, are not prevented from voluntarily disclosing certain information about research subjects, such as evidence of child abuse or a subject's threatened violence to self or others. However, if a researcher intends to make such voluntary disclosures, the consent form should clearly indicate this.

The HRS has a CoC, providing additional confirmation that the project would be of appropriate scope to request one. Discussion with HRS advisors seems sensible, and additional detailed information may also be found at:
http://grants.nih.gov/grants/policy/coc/index.htm

4.2 Data Sharing

A primary goal of the Kavli HUMAN Project is to make the data collected under its auspices an open resource for scholars, but balancing data sharing with privacy concerns is not trivial. Sharing with researchers from industry, such as pharmaceutical

company employees has additional potential for concern, particularly since it is critical to maintain independence from for-profit entities. Therefore, it will be important to ensure that strong governance policies are in place to guide decisions about data sharing.

Policies should be as permissive as possible, but the beneficial outcomes of analyses must be weighed against ethical concerns, de-identification risks, the possibility of adverse legal bias or oppression and potential public relations ramifications, fallout from which could compromise the feasibility of the project by compromising relations with participants and funders. Each research proposal that requests use of the data will be evaluated with respect to these criteria - articulated in the form of data governance policies – by a committee including representatives from all of the stakeholder communities, the *Research Proposal Evaluation Committee*. This will include civil rights advocates, privacy experts, study participants, academic researchers and corporate researchers. An oversight monitoring and auditing mechanism will also be necessary.

At this time, the Privacy and Security Advisory Council has advised that we not allow use of the data for commercial or marketing purposes, but there may be industry users who may have substantial interest in using the data for health or education research. There is wide agreement between the Council and the KHP Team that such a use falls well within the study's overall mandate. Nonetheless, as the research results would benefit for-profit corporations, it may make sense to charge these entities for use of the data. A tiered pricing model would take into account a number of user factors, for example the size of the organization and the number of data points or data stores being accessed. There are some existing licensing models of this type that we might look to for guidance. However, profit or commercialization of the database are not study goals, and would certainly have negative ramifications. Therefore, it will be critical to make clear that any money obtained from for-profit users will be reinvested in providing financial aid to researchers from non-profit institutions. This is also important because if the project does make a profit, there may be issues of property rights – i.e. what people should get out of their own data if we are profiting from it. During the CD2 phase a proposed fee structure and regulations for users of different types of data will need to be drafted.

4.3 Recombination of Our Data with Administrative Data

One way of increasing the power of our data would be to combine it with government records from the Census Bureau, the IRS and the Social Security Administration, as well as data from social programs administered by the state government, such as SNAP, Medicare and TANF. However, each of these entities has their own privacy regulations, so it will be necessary to comply with all relevant policies, particularly with regard to data recombination. As an aside, it is relatively easy for individuals to get their own SSA and IRS records released to them, but the protocols typically require that the data be sent to the individual, not to a third party. Some discussion with the HRS leadership may be necessary to determine whether it is possible to arrange bulk release of records.

However, recombination of our data with detailed census data will require a unique protocol, as such data is only accessible at a Census Research Data Center. There is already an RDC in Manhattan, at Baruch, though it might be possible to create one at NYU/CUSP. However, in either location, identified data would have to be extracted from our database to combine with census data on the RDC servers. Consequently, we would need to get consent from participants to give our identified data (including social security numbers) to the government for linking. During the CD2 and/or CD3 process, we will carefully consider opportunities for recombination and outline a protocol for implementing such analyses while ensuring the security and privacy of our participants' data.

5. Consent

5.1 The Main Consent Process

Getting informed consent for this complex study will require a process carefully designed to ensure that all participants understand the risks and benefits of the research we hope to undertake. Given that there are analyses that we have not yet conceived (with equally unknown outcomes) it will require a careful balance between being as specific as possible and leaving open future possibilities. We expect that there will need to be additional consent processes as the study progresses, but we plan for an initial comprehensive consent process that will address the first five years of the study.

We envision this main consent process as a three-stage sequence. First, a video will be presented which describes the basic issues that require consent, as well as insights into secondary use cases for the data. The video will be divided into segments, and after each one, a member of the consent staff will stop the presentation to administer comprehension checks and facilitate a discussion about the issues raised in the video. After the video, participants will be presented with the opportunity to provide oral consent. If they do so, the third step will be the presentation of the long paper form for participants to sign. This paper form will contain all of the legal language necessary for written consent, but all the information in the form will have been covered during the video and discussion process. Thus, we expect that the acquisition of written consent will be relatively straightforward and a somewhat incremental step of the consent process. The intake process will be scheduled no sooner than twenty-four hours after the consent process, so that participants will have time to re-consider their participation before data collection begins.

There are a number of ways that the full consent process could be administered. One possibility would be to perform it as the first step of the intake process. However, this has a number of disadvantages – most importantly, it does not allow much of a waiting period before data collection begins, and logistically it would make it more difficult to efficiently schedule data collection during the intake process, since in order to enroll all members of a family, they would have to spend the first part of the day participating in the consent process. Instead, we propose that the consent process be administered in the participant's own home a day or two prior to the start of the intake process. This may also help to put people at ease, as they will be in a familiar space. Estimated staffing costs for this process are described in the preliminary budget analysis (Appendix B).

5.2 Additional Consent Processes

The consent process will be administered for all members of the residential network at the same time, prior to intake. However, there may also be a need to get consent from people who are not members of the study. Non-residential family members who we hope to follow as auxiliary members of the study (doing less intensive collection of genetic and behavioral data) will need a separate consent process, and it may also be appropriate to obtain consent from care providers who spend substantial amounts of time in the home (i.e. nannies and home health care aides). As we plan to collect only metadata about communications partners, it has been tentatively concluded that it is not necessary to obtain consent from them, but as there may be a privacy issue (or the perception of one), it may make sense to provide some notice of disclosure during the communications to our participants.

We expect that the initial consent process will provide adequate protection for study participants in achieving the main study aims. However, over time, it may be advantageous to expand the scope of the study with the use of additional technology or secondary use cases. These situations will likely require additional consent from our participants. As described previously in this document, for the preservation of privacy, it will be necessary for the study staff to handle this process. For efficiency, it would be preferable if additional consent occurred at the biannual appointment for physical specimen

collection, phasing in large scale changes over time, but an additional appointment with study staff could be arranged for urgent projects or for those based on a small subsample of the subject pool.

5.3 Ensuring the Protection of Children in the Study

A major challenge for ensuring informed consent in study participants is the collection of detailed data (particularly genetic data) from children. As required by law, parents will provide consent for their children to participate in the study and children will assent to their participation. At NYU, children twelve years or older are considered capable of providing written assent. When participants reach age eighteen, they will have to go through the consent process as adults and agree to continue to participate in the study. Given that the research here will not provide a direct benefit to participants, we note that it was the consensus of the Council that no child can be compelled to provide data of any type unless he or she freely assents. This is true even if the child's parent or guardian assents. So, for example, a 9 year-old child cannot be compelled by his/her parents to permit a blood draw for the purposes of the study. *This is a critically important limitation on data collection from children.* We therefore assume that invasive measurements such as a blood draw may not be possible in children or teenagers. However, we do expect to be able to perform genetic sequencing (which does not require blood), as well as detailed educational and cognitive profiling. These are measurements that will result in the collection of extremely sensitive data about children. While parents may feel comfortable consenting for their children to contribute such data to the project, young children may not have the capacity to understand the implications of providing genetic data or long-term detailed data collection in general. For this reason it remains undecided whether it may be more ethically justified to give participants reaching adulthood not just the opportunity to leave the study, but also the opportunity to withdraw their highly sensitive data from the database – even if that data has been gathered over 18 preceding years. Further discussion of the ethical obligations and consideration of the policies of other longitudinal studies (such as the NLS and NCS) will be necessary before making a final decision on this issue. Assessing these policies will need to be completed before the end of the CD2 phase.

Another issue for discussion is whether the Children's Online Privacy Protection Act (COPPA) is relevant to the study. COPPA pertains to websites collecting information from children under the age of 13. If a web-based authentication process is used for data collection, COPPA may apply. This regulation was issued by the FTC, and is largely about obtaining parental consent for any data collection, which is already incorporated into our procedures. However, it does entitle parents to request deletion of their child's data at any time, which could have serious consequences for the study. Further legal advice on this matter will be necessary to determine the implications for study policies prior to the completion of the CD2 phase.

5.4 Other Vulnerable Populations

Over the course of the study, we expect that some elders will begin to lose their mental faculties. For those that are legally declared incompetent, there will be an individual designated as having power of attorney, and this agent will be the person responsible for any decisions or consent requirements. However, there may be situations where study staff, through interactions with a participant, may perceive that mental function is compromised, though no legal measures have been taken. In these cases, a member of the study staff will need to identify an advocate or interested agent with whom they can consult. However, some policy will need to be set in order to determine the threshold for invoking this process. These policies will need to be defined and legally vetted prior to completion of the CD3 stage.

Over the course of the study, some participants will also go into prison, where active data collection will be difficult. However, it may still be possible to collect data passively (especially court and arrest

data, which are a matter of public record), and certainly we would hope to be able to use any data on these people to understand the role of prior experience in post-incarceration outcomes. To that end, it will be necessary to put in place policies to address these issues, and to justify the importance of this research as it benefits the prison population as a whole, and more generally supports studies of recidivism.

6. Conclusion

The success of the Kavli HUMAN Project depends critically on protecting the valuable resources that will be developed in the course of the study – the rich set of data as well as the privacy and trust of the participant population. The data governance and security measures described in this section outline the methods and approaches that will be used to achieve these aims. In the next stage of study design, we will specify the procedures and policies that will ultimately be implemented.

6

EDUCATION AND PUBLIC OUTREACH

1. Introduction

The Kavli HUMAN Project (KHP) is a first-of-its-kind study of human beings, and promises to deliver previously unachievable insights into human health and behavior, new therapeutics, and recommendations for evidence-based public policies. The KHP has the potential to be the deepest and most impactful novel ever written about the human condition. Rarely has a scientific study possessed the potential to improve the lives of so many people on so many different fronts in specific and measurable ways. But for the KHP to reach its potential, open lines of communication must be fostered with all of the project's stakeholders – scholars, potential participants, and potential funders in both the foundation and government communities.

Communicating the importance and benefits of basic research is crucial to fostering sustainable long-term support amongst the public for research projects like the Kavli HUMAN Project. The KHP team understands the necessity of a robust education and public outreach (EPO) strategy and is creating an EPO infrastructure to oversee and implement this strategy. The KHP team also recognizes the communication challenges that await, given the nature of the Kavli HUMAN Project's data collection and surrounding issues of privacy and data security. Nevertheless, as this section of the Preliminary Study Design Report explains, the KHP team is excited and ready to tell the story of the Project and communicate its promise, but is also prepared to address the comments of even its harshest critics.

A smart narrative offers the best chance to break through the constant bombardment of news and social media that overwhelms the average person in order to grow the visibility of a project like this. Beyond that, any effort—scientific or otherwise—that involves solicitation, cataloging, and sharing of personal data must show itself to be cognizant of people's fear about sharing that information. Addressing these concerns up front—by proactively highlighting the extensive security procedures being put in place, for example—will be critical.

Another critical element of our communications strategy involves the KHP Education and Public Outreach Advisory Council (EPOAC), the membership of which will consist of experts in different medical and science fields, as well as in public relations and science communication. Our EPOAC members will advise on the KHP's outreach strategy, use their stature and connections to publicly advocate for and advance the reputation of the KHP, and serve as a bulwark against negative criticism.

Failure to communicate effectively to both the project's key stakeholders and the public about the benefits and motivations of the Kavli HUMAN Project risks dooming the effort before it even gets underway. Overcoming a negative public perception is far more difficult than preemptively building a positive one from scratch. Letting any negative definition take hold will also embolden the inevitable opposition to any project of this nature and size.

This section of the Preliminary Design provides a detailed overview of the communications environment the KHP is being launched in; the education and public outreach strategy and constituent elements thereof; the messaging and audiences with whom we will be engaging; events planned over the course of the next two years leading up to the beginning of enrollment; and the resources necessary to fulfill this overall strategy.

2. Big Science in a Changing Media Landscape

The Kavli HUMAN Project is being launched into a media and public discourse environment that has transformed dramatically in only the past few years and shows no signs of abating. This change is compounded by other cultural changes related to the public's perception of science, scientists, and institutions more generally. The scientific community is under constantly increasing pressure to engage with the public and explain the nature, purpose, and—more often than not—the benefits of their research. In a constricting fiscal environment for public funding of scientific research, research increasingly needs to be funded from multiple sources, public and private. More pressure on funding sources, issues with public scientific literacy, and other forces acting upon the scientific community have led to the burgeoning field of science communication, once left to the realm of a few elite journalists and scientific publications. If the Kavli HUMAN Project is to be successful, that success will in large part depend on its communication efforts with the academic community, local, state and federal actors, press, and the New York City population that KHP will study. Understanding the current media environment is crucial for identifying the methods and tools required for a successful communications campaign.

Knowledge today is disseminated very differently from a generation ago, with the lines between science, the media, and politics becoming increasingly blurred. Today's information for public consumption, including scientific research, is increasingly mediated for the public, with the issues of the day being constantly dissected through different ideopolitical lenses before ever reaching the average person.[3] This trend is compounded by the specialization of media outlets in terms of coverage topics and/or ideology, and individuals' narrowing spectrum of content curation to consume news media that predominantly reflects their own views and interests. Despite the generally high-quality scientific reporting and coverage in major traditional media outlets like *The New York Times* or *Scientific American*, Americans are increasingly getting their news from alternative sources and staying within their content "bubble."[4] A recent report suggests that the number of Americans following science news "very closely" has decreased over the past two decades.[5]

One of the advantages of this changing media landscape, however, is the proliferation of science and technology-oriented outlets and content across all media platforms. From NPR's Science Friday to the Discovery Channel to Gizmodo.com, science has entered mainstream pop culture and the nerd is cool once more.[6] Furthermore, the Internet allows anyone to communicate directly with the public and any other stakeholders through online web and social media platforms, whereas before, people wishing to communicate with a large audience were beholden to media interests.

---

[3] Scheufele, D.A. (2014). Science communication as political communication. *Proceedings of the National Academies of Sciences,* Vol. 4, Suppl. 4. http://www.pnas.org/content/111/Supplement_4/13585.full.pdf

[4] Clough, G.W. (2011). *Increasing Scientific Literacy: A Shared Responsibility*. Smithsonian Institution. Washington, DC. http://www.si.edu/Content/Pdf/About/Secretary/Increasing-Scientific-Literacy-a-Shared-Responsibility.pdf

[5] Scheufele, D.A. (2014). Science communication as political communication. *Proceedings of the National Academies of Sciences,* Vol. 4, Suppl. 4. http://www.pnas.org/content/111/Supplement_4/13585.full.pdf

[6] Harrison, A. (2013, September 2). Rise of the new geeks: how the outsiders won. *The Guardian*. http://www.theguardian.com/fashion/2013/sep/02/rise-geeks-outsiders-superhero-movies-dork

Science celebrities are also becoming increasingly visible, serving as advocates for science, technology, education, and mathematics (STEM) education and investments in research, as well as providing a reasoned viewpoint on the science debates of the day. NASA, for example, brought in Will.I.Am, front man of the popular band, Black Eyed Peas, to help drum up interest in their Mars program[7], and continues to work with other celebrities. Yet, despite their inroads, even the respectable 7 million combined Twitter followers of celebrated science communicators like Neil deGrasse Tyson and Bill Nye "The Science Guy" pales in comparison to the 67 million (approximately the population of France[8]) who follow Justin Bieber. However, the trend generally remains positive.

Despite the rise of interest in STEM topics and coverage thereof, faulty claims and myths are still frequently shared as fact on social media sites.[9] These competing trends suggest that while the cultural landscape is becoming friendlier to STEM and those who pursue careers in science, science literacy in the United States has not kept pace with the rate of discoveries and technological advancements, and indeed appears to have eroded. A 2009 national survey of the California Academy of Science found that,

> "…only 59 percent of adults knew that early humans did not coexist with dinosaurs; only 53 percent knew how long it takes the Earth to orbit the sun; only 47 percent could give an approximation of how much of the Earth's surface is covered with water; and only 21 percent knew all three of these things."[10]

While no one can be an expert on everything, a declining understanding of the basic precepts of science, like the scientific method, is muddling coverage of scientific breakthroughs and disparaging areas of research or clinical use. The KHP can look at other large scientific endeavors for lessons on how to approach the aforementioned challenges, while also highlighting differences between the KHP and those initiatives that the Project's education and public outreach strategy must take into account in order to ensure success.

Historically, large-scale science has had somewhat limited interactions with the media. However, the Kavli HUMAN Project differs greatly from other large modern science endeavors, like the Sloan Digital Sky Survey (SDSS), U.S. Health and Retirement Study (HRS), or Large Hadron Collider (LHC). Only one of those projects, the HRS, dealt with human subjects in a very focused way: understanding how the elderly age and make decisions, and what potential remedies for different ailments are possible based on their research. In many respects their goals are similar to the KHP, though more limited in scope. What differs even more between the two studies is that the HRS relies primarily on interviews, surveys, and traditional social science data collection methodologies. The KHP's data collection process will be considerably more all-encompassing, and falls squarely in the current societal conversation about privacy and data security.

In a completely different area of scientific research, the Large Hadron Collider is plumbing the depths of theoretical physics, but was the only of these activities to receive any appreciable negative press from unfounded theories that the Collider would create a black hole that would destroy the Earth. The

---

[7] Trotta, A.M. (2012, August 28). *Curiosity Rover Plays First Song Transmitted From Another Planet*. NASA. http://www.nasa.gov/mission_pages/msl/news/msl20120828.html

[8] *Demography: Population at the beginning of the month, France except Mayottte* (2015, September 9). National Institute of Statistics and Economic Studying. http://www.insee.fr/en/bases-de-donnees/bsweb/serie.asp?idbank=001641607

[9] Johnson, G. (2015, August 24). The Widening World of Hand-Picked Truths. *New York Times*. http://www.nytimes.com/2015/08/25/science/the-widening-world-of-hand-picked-truths.html

[10] Clough, G.W. (2011). *Increasing Scientific Literacy: A Shared Responsibility*. Smithsonian Institution. Washington, DC. http://www.si.edu/Content/Pdf/About/Secretary/Increasing-Scientific-Literacy-a-Shared-Responsibility.pdf

LHC, along with the entire physics community, provided a forceful rebuke of the faulty claims, in the form of articles in the popular press from leading researchers[11] and detailed press releases[12] from the organization that operates the Large Hadron Collider, CERN, which similarly engaged with and quoted renowned scientists to help combat the myth to much success. The KHP can learn from this and will employ similar measures should the need arise, but what separates the KHP from these and other large-scale scientific endeavors is that we will not wait to be placed in such a situation to deal with it, but rather are building in the risk of negative coverage and public reaction into our overall development and implementation strategy to minimize risks. A comprehensive crisis communication plan will be built into our overall communications strategy, and elements of that plan are detailed further in this section.

However, neither the SDSS nor the LHC deal with people as the focus of their research, nor does their research deal with the very sensitive information collected by studies of people. To succeed, among the many other domains listed in this Preliminary Design document, the Kavli HUMAN Project will need to play a proactive role in shaping the narrative surrounding its mission, goals, and methods. In an era where "Big Brother" is a familiar refrain and boogeyman, and now quite achievable thanks to advances in technology, the KHP team is keenly aware of the critical ammunition that could potentially be used against the Project, and will be very open about this potential route of criticism. However, the altruistic promise of the Kavli HUMAN Project holds an appeal of its own, and when combined with the stringent cybersecurity and data governance measures built into the Project, will be able to overpower negative criticism.

[11] Lincoln, D. (2015, March 25). The Truth About Black Holes, the Large Hadron Collider, and Finding Parallel Universes. *The Huffington Post*. http://www.huffingtonpost.com/don-lincoln/misleading-science-black-_b_6934296.html

[12] *The Safety of the LHC*. European Organization for Nuclear Research (CERN). Accessed September 14, 2015. ttp://press.web.cern.ch/backgrounders/safety-lhc

The KHP will operate in a rapidly evolving landscape of science communication and perspectives. To successfully navigate this space, the Project will take advantage of the many new communication channels available to reach out to our diverse set of stakeholders. At the same time, it will be critical to address any misconceptions about the KHP, from the potential science conducted to the public policy implications, which impede understanding of the Project. The KHP team will remain vigilant in explaining the aspirational rationales for conducting a project of this scope, the societal benefits of the KHP, and actively engaging with a diverse set of stakeholders before, during, and after enrollment of the Project cohort.

The challenges presented to the KHP by a changing media landscape and science's evolving role in the zeitgeist bring us to one conclusion: **Taking education and public outreach efforts for granted will make it impossible for the KHP to achieve its aims.**

3. Messaging and Best Practices for the Kavli HUMAN Project

Because of the diverse stakeholders involved and reliance on the public for accomplishing the goals of the Kavli HUMAN Project, conveying the promise and nature of the KHP will be a core activity of the Project, and instrumental to its success. This necessitates recognizing who our different audiences are, honing the right messages for those audiences, and conducting research to determine what people are most receptive to or negative towards. In addition, it is not only important to know what to say, but to identify best practices for communication (i.e. proactive versus defensive messaging). We can also draw lessons learned from other organizations that have dealt with messaging challenges to improve our communications preparedness and overall success.

We consider the KHP to be a "moonshot" project, one of massive scale that will spur major advances in the understanding of human beings. In addition

to associating the KHP with other transformative science and technology initiatives, long-term outreach campaigns will also highlight the individuals and families who make up the KHP study population to show the diversity of our participants and engender positive feelings towards the Project. From a messaging standpoint, there are many different positive aspects of the KHP that we can emphasize, for instance harnessing the power of measurement for better science and for the greater good, and working across disciplines to achieve discoveries that would otherwise be out of reach.

At its heart, perhaps *the* overarching messages for the Kavli Human Project is that it seeks to study human beings in an unprecedented manner to fully understand the relationship between biology, behavior, and the environment—something unachievable with current research methods. It is a platform to obtain new measurements that advance the human condition at both the micro and macro levels, from better understanding everyday decisions to elevating the practicality of public policy in the United States. The goal is to develop real-world comprehension of the roots of human behavior and apply that to fundamentally improve quality of life.

When it comes to measurement, we can explain the significant challenges to scientific progress that arise when one's instruments reach their limits. In a 2013 piece for *The Wall Street Journal*, Bill Gates wrote: "I have been struck by how important measurement is to improving the human condition. You can achieve incredible progress if you set a clear goal and find a measure that will drive progress toward that goal."[13] There is a shared belief in the scientific community that, more often than not, scientific revolutions are driven by changes in measurement rather than changes in theory. Even some of the most impactful of longitudinal studies—the U.S. Health & Retirement Study (HRS), the Framingham Heart Study, the Nurses' Health Study (NHS)—have been limited in depth.

Regarding working across disciplines, an NIH-funded Institute of Medicine report noted the "traditional and persistent barriers to interdisciplinary research…The barriers might best be presented in five major categories: attitude, communication, academic structure, funding, and career development. Despite the hesitation of some scientists to engage in interdisciplinary research, the nature of the complex scientific challenges that we face creates a need to ensure that it can occur."[14] The Kavli HUMAN Project will not only encourage, but also enable, interdisciplinary research. With initial research outputs possible just 3-5 years after enrollment and measurement commence, the interdisciplinary data will be turned into research results in a relatively short turnaround; for example, cross-correlating environmental data, like toxin exposure, with epigenetic mutations or changes in brain structures gathered through genome sequencing and MRI imagery to identify what environment factors affect changes in our biological and behavioral processes.

In addition to this example, a diverse set of use cases for the KHP data are described in a group of White Papers commissioned by the Scientific Agenda Advisory Council (described in more detail in chapter 7 on the scientific agenda). These research questions to be explored using the KHP dataset demonstrate that the Project will enable researchers to conduct impactful science that can improve people's lives, from providing real-life context for laboratory-derived neuroscientific data to quantitatively mapping phenotypes to enable precision medicine therapeutics, there are many examples. Specific research identified by the Scientific Agenda Advisory Council for exploration include:

[13] Gates, B. (2013, January 25). Bill Gates: My Plan to Fix the World's Biggest Problems. *The Wall Street Journal*. http://www.wsj.com/articles/SB10001424127887323539804578261780648285770

[14] Committee on Building Bridges in the Brain, Behavioral, and Clinical Sciences (2000). *Bridging Disciplines in the Brain, Behavioral, and Clinical Sciences*. Washington, DC: National Academies Press. http://www.nap.edu/catalog/9942/bridging-disciplines-in-the-brain-behavioral-and-clinical-sciences

- **Rewards and Economic Decisions: From Lab to Field**
    - Neuroscientific research has traditionally been limited to laboratory settings, precluding many kinds of research that could help advance, and even revolutionize, the field. The Kavli HUMAN Project will allow researchers to determine real-world applications of laboratory work, in particular work on the dopamine system and reward prediction errors, and our mechanisms for self-control. KHP will help identify circumstances under which reward signals overcome self-control mechanisms, as may occur in such cases as drug addiction and diet. KHP will enrich our understanding of how environmental conditions and differences across individuals impact the balance between the pull of reward and opposing self-control mechanisms.
    (Wolfram Schutz, Wellcome Principal Research Fellow, Prof. of Neuroscience, Univ. of Cambridge; Fellow, Royal Society)

- **Longitudinal Studies in Neuroscience: Understanding Human Dynamics**
    - Psychology and neuroscience currently have almost zero knowledge about how the mind and brain change over time scales of days, weeks and months. This is particularly problematic given that most major mental health disorders (including depression, schizophrenia, and bipolar disorder) involve drastic fluctuations in mental function over the course of weeks. The Kavli HUMAN Project will allow researchers to examine changes in behavior and mental function over time scales of days, weeks, months and even years. Such data will provide the opportunity to link these relatively rapid changes to mental health disorders, as well as to provide new insights into existing research on slow changes previously identified in studies of child development and aging.
    (Russell Poldrack, Prof. of Neuroscience and Cognitive Psychology, Stanford University; Member, Stanford Neurosciences Institute)

- **Life Adversity during Sensitive Periods of Brain Development: Social & Economic Impact**
    - Repeated trauma and abnormally high, prolonged stressful experiences—particularly adversity from caregivers—target the development of emotional and cognitive circuits in the brain. Environmental neurotoxins like pollution can amplify these effects. Recent work suggests that social support, particularly from caregivers, can buffer the effects of adversity on the developing child's brain. Capitalizing on new technology, the KHP will assess the amount and quality of social contact between the child and others in the honeurome environment, suggesting potential mechanisms by which the social interactions can influence childhood brain development.
    (Regina Sullivan, Prof. of Child and Adolescent Psychiatry, New York University School of Medicine; Research Scientist at Nathan S. Kline Institute for Psychiatric Research)

- **Real-Time Assessment of Wellness and Disease in Daily Life**
    - Biological markers have been directly associated with disease risks, but poor measurement of behaviors such as diet and exercise limit our understanding of preventive measurements. By joining together an uncommonly wide range of disciplines and expertise, the Kavli HUMAN Project will advance measurement of behavioral phenotypes, as well as environmental factors that impact behavior. The longitudinal nature of KHP will liberate new

understanding of dynamic links between behavioral phenotypes, disease, and the broader environment and advance understanding of the bio-behavioral complex. The combination of measurements will seed new approaches to the diagnosis, prevention, and treatment of human disease.

(Dennis Ausiello, Jackson Distinguished Prof. of Clinical Medicine and Emeritus Physician-in-Chief of Harvard Medical School; Co-Founder & Director, Center for Assessment Technology and Continuous Health (CATCH), Massachusetts General Hospital; and Scott Lipnick, Scientific Director, Center for Assessment Technology and Continuous Health (CATCH), Massachusetts General Hospital)

- **How Genetic and Other Biological Factors Interact with Smoking Decisions**
  - Associations have been established (or at least, widely accepted) between smoking and genetics, but there has been little success in determining what other factors lead to behavior that negatively impact health. The KHP will focus on the lifecycle trajectory of smoking as well as interactions with other choices and environmental factors to enable the design of more effective public policy.

    (Laura Bierut, Co-Director, Outpatient Clinic, Washington Univ. in St. Louis School of Medicine and David Cesarini, Asst. Prof. of Economics, Center for Experimental Social Science, NYU)

- **Long-Term Care in the Family Context**
  - Families of all social, economic and cultural backgrounds struggle with the best way to handle the long-term care needs of older loved ones. The emotional and financial burdens associated with long-term care can put tremendous stress on everyone involved. The KHP will look at how these decisions impact both the recipients of care and those who provide it. The resulting data will help families prepare for long-term care decisions, improve public policies and inform discussions of insurance options.

    (Andrew Caplin, Silver Prof. of Economics, NYU and Kathleen McGarry, Chair, Department of Economics, UCLA)

Messaging will also be tailored by audience while still identifying cross-cutting themes and messages that can be used more widely. There are two types of messaging approaches that we will employ: primarily we will use *proactive* messaging to explain the positive, beneficial nature of the KHP, and *defensive* messaging will be used when needed to correct mischaracterizations of or attacks on the Kavli HUMAN Project.

3.1 Proactive Messaging

An analysis of several data-driven research projects suggests that the three most important proactive messaging strategies for scientific projects of this nature are to (1) define the need; (2) define the goal; and (3) address concerns proactively. KHP leadership and existing literature already encompass these best practices to some extent, but previous major research projects like the Human Genome Project and Sloan Digital Sky Survey provide analogs for the Kavli HUMAN Project to learn from for each of these three proactive messaging domains:

Define the need:

In the 1970s, the U.S. scientific community responded to the challenge of a fragmented genetic community by proposing and executing the Human Genome Project (HGP). Rather than a piecewise approach to understanding genetics, a group of visionary scientists proposed a complete catalog of the human genome – a proposal to transform genetics from a series of isolated fiefdoms into a global synthetic field.

The leaders of the Human Genome Project (HGP) spoke frankly about the need for the project. Dr. Charles Cantor, one of two HGP directors, said, "the problem up to now was that gene mappers have been using four or five different techniques to identify clones. The only way to study a marker discovered by another scientist was to get a sample of the genetic material. These samples were difficult to handle, and competitive researchers were often unwilling to share them."[15]

HGP's importance in the field of research is impossible to overstate. The Project's own website makes the following analogy: in 1911, Alfred Sturtevant, then an undergraduate researcher in the laboratory of Thomas Hunt Morgan, realized that he could - and had to, in order to manage his data - map the locations of the fruit fly (Drosophila melanogaster) genes whose mutations the Morgan laboratory was tracking over generations. Sturtevant's very first gene map can be likened to the Wright brothers' first flight at Kitty Hawk. In turn, the Human Genome Project can be compared to the Apollo program bringing humanity to the moon.[16]

The resulting accomplishments, including the sequencing of the human genome, revolutionized the biological sciences.

For the field of astronomy, the power of this comprehensive approach can also be seen in the Sloan Digital Sky Survey (SDSS). Prior to the development of the SDSS, data collection in astronomy was dominated by individual astrophysicists competing for limited access to telescopes, which they used to collect limited data sets that were primarily useful for addressing a specific question. The SDSS revolutionized the field by shifting to a new approach – a systematic examination of the sky. The most efficient instruments were used to build a comprehensive, publicly-available dataset applicable to a broad range of questions in astrophysics. This resource has become invaluable, supporting thousands of publications by astrophysicists and researchers around the world.

The KHP needs to do for social science what the HGP and SDSS did for biochemistry and astronomy, respectively, but first has to explain why it is the way to move forward to advance so many fields of human health and behavior. The KHP will "define the need" by contrasting the KHP with other studies of human beings and explaining their inherent limitations that can only be overcome by a study of the Kavli HUMAN Project's nature.

Define the goal:

Projects of this nature need to simply and succinctly justify the need for their own existence. In 1989, as scientists prepared to launch the HGP the following year, they put forth very specific goals and needs. "The goal is an understanding of how the healthy human body works and how the genes are involved when the body is sick,"[17] wrote *Newsday.*
HGP and other studies have since advanced our understanding of the healthy human body immeasurably. The problem though, is that many of our tendencies – the decisions we make, the actions we take – have not caught up. The goal of the KHP then is to bridge that gap between staggering advancements in our understanding of biological systems and the lesser-understood forces that shape our decision making and behavior.

[15] Blakeslee, S. (1989, October 10). A Roadmap for Genes. *The New York Times.* http://www.nytimes.com/1989/10/10/science/a-road-map-for-genes.html

[16] *What is the Human Genome Project?* National Human Genome Research Initiative. Accessed on September 14, 2015. https://www.genome.gov/11511417

[17] Cook, R. (1989, July 25). Scientists hope to take advantage of new human-gene swapping tools to genetically engineer a host of animals to produce more milk, meat, eggs and wool. *Newsday.*

Address the Concerns Proactively:

One of the most obvious concerns about the KHP will be the safeguards in place to protect participants' data. This is a common worry about studies of this nature and the best practice is to address it proactively and directly. The Health of Women (HOW) Study is one of several that devotes an entire section of the project's website to communicating an acute awareness about the issue, and laying out specific policies from the beginning. "The HOW Study … considers it our utmost responsibility to safeguard your privacy,"[18] notes the site. It also lists with granularity some of the specific security measures put in place: offsite secure hosting environments, encryption technology, etc.

3.2 Defensive Messaging

Even the most well-intentioned security protocols sometimes fail to prevent a data breach and the crisis communications plans enacted in response can determine the extent of brand damage suffered. The case studies below offer lessons that the KHP must incorporate into its plans regardless of a future problem being related to an actual data security breach, issues pertaining to study participant privacy, or even just the *perception* of a problem in either case:

Example 1, Data Breach: *Anthem Health Insurance*: The largest such breach in the health space was the attack on Anthem, the nation's second-largest health insurer, in late January 2015. Eighty-million customers and employees had their personal data compromised when hackers gained access to Anthem's network. The company was quick to acknowledge the breach, bringing the information directly to authorities and communicating it to customers both directly and indirectly (via the media). Anthem publicly announced the breach within a few days of its discovery (despite the fact that federal regulations permitted them to wait longer). Anthem's C.E.O. apologized directly to customers and the company quickly provided frank and concise communications—via both its website and regular mail—about what customers needed to know. Anthem also voluntarily provided free credit monitoring to anyone whose data was compromised.

Home Depot and Target: Both companies suffered massive cyber-attacks in 2013 and 2014, respectively. In both cases, hackers obtained the personal information of tens of millions of customers; in the Target case, financial data was also stolen via credit card transaction records. Both companies took immediate and public steps to mitigate the brand damage, including both direct-to-customer and media communication, full and public cooperation with investigating authorities, amends to affected customers and, in Home Depot's case, changes in senior company leadership. As a well-regarded example of direct-to-consumer communication, here is the 2014 letter that Gregg Steinhafel, the president, chairman, and C.E.O. of Target, sent to customers[19]:

> *Dear Target Guests,*
>
> *As you have probably heard, Target learned in mid-December that criminals forced their way into our systems, gaining access to guest credit and debit card information. As a part of the ongoing forensic investigation, it was determined last week that certain guest information, including names, mailing addresses, phone numbers or email addresses, was also taken.*
>
> *Our top priority is taking care of you and helping you feel confident about shopping at Target, and it is our responsibility to protect your information when you shop with us.*

[18] *Privacy and Consent*. The Health of Women Study. Accessed September 14, 2015. https://www.healthofwomenstudy.org/privacyandconsent.aspx

[19] Steinhafel, G. (2014). Letter from Chairman, President, and Chief Executive Officer of Target to the Public. Accessed September 14, 2015: https://corporate.target.com/_media/TargetCorp/global/PDF/GreggLetter-ad-version04.pdf.

We didn't live up to that responsibility, and I am truly sorry.

Please know we moved as swiftly as we could to address the problem once it became known, and that we are actively taking steps to respond to your concerns and guard against something like this happening again. Specifically, we have:

1. Closed the access point that the criminals used and removed the malware they left behind.
2. Hired a team of data security experts to investigate how this happened. That effort is ongoing and we are working closely with law enforcement.
3. Communicated that our guests will have zero liability for any fraudulent charges arising from the breach.
4. Offered one year of free credit monitoring and identity theft protection to all Target guests so you can have peace of mind.

In the days ahead, Target will announce a coalition to help educate the public on the dangers of consumer scams. We will also accelerate the conversation–among customers, retailers, the financial community, regulators and others–on adopting newer, more secure technologies that protect consumers.

I know this breach has had a real impact on you, creating a great deal of confusion and frustration. I share those feelings. You expect more from us and deserve better.

We want to earn back your trust and confidence and ensure that we deliver the Target experience you know and love.

We are determined to make things right, and we will.

Example 2, Privacy Misperceptions: *Titan:* A data breach is not necessary to create a communications crisis. For Titan, a Manhattan-based advertising company, simply the perception that data was being gathered without the public's permission or knowledge was enough to cause days of negative headlines. In October 2014, BuzzFeed reported that Titan, which controls advertising on New York City's payphones, had installed thousands of *"tiny radio transmitters known as 'beacons' — devices that can be used to track people's movements — in hundreds of payphone booths in Manhattan, BuzzFeed News has learned. And it's all with the blessing of a city agency — but without any public notice, consultation or approval."*

The seemingly stunning news generated more than 100 news stories and days of coverage from tabloids, network news, and tech reporters across the country. In fact, the technology was only being tested, not put into actual use, and beacons are an opt-in technology that mobile phone users must specifically allow before they can track a user's location. The City and Titan tried to correct the record, but once the story had momentum, the facts became irrelevant. The media narrative was that something that sounded scary was being done with no public notice or debate. Titan was forced to cancel the testing and suffered lasting damage to its brand.

This incident speaks to how sensitive media, especially in New York City, can be about both privacy and personal data. It also demonstrates the need to be proactive and fully transparent about KHP practices and policies, as things that may seem innocuous to scientists could feel threatening to participants and the public if not fully disclosed and properly explained.

There are a number of components to any successful rapid response plan centered around a data breach or participant privacy: predetermined internal and external protocols, including a war room; proactively informing both authorities and stakeholders about the issue; and concrete steps to help anyone affected manage the damage. Considerations for a crisis communications plan are detailed further in this section.

4. Core Audiences and Associated Messaging

4.1 Core Audiences

The Kavli HUMAN Project will depend upon, and need to maintain relations with, a wide range of stakeholders, both in the lead-up to enrollment of the first study volunteers and beyond. The KHP team has identified the following as our main audiences, and has tailored its event and messaging strategies to each audience, while accommodating for overlap:

- Academia
- Media
    - Local & National
    - Academic & Popular
- Policymakers
    - Local & State
    - Federal Science Agencies
    - U.S. Congress
    - Executive Office of the President
        - Office of Science and Technology Policy
        - Office of Management and Budget
- General Public
    - New York City residents (the participant pool)
    - National Public

Figure 6-1 shows the communications landscape the Kavli HUMAN Project is entering and the interplay between stakeholders and the KHP. Although our ultimate communications strategy will be tailored for each entity, the figure below shows that while it is relatively easy for the KHP to engage with academia and policymakers without intermediaries, direct public engagement is a more difficult matter, as most of their content comes from media sources, not directly from academicians or policymakers. The weight of the lines connotes their relative importance, with the heaviest connection being between the media and public, signifying the necessity on the Kavli HUMAN Project's part to build and maintain strong and active ties with media outlets. However, the weight the media carries can be leveraged, and when necessary overcome, through a concerted outreach effort, as described later in this section. Implicit in this diagram is that the KHP is open to feedback from all stakeholders at all times to strengthen the study design and implementation.

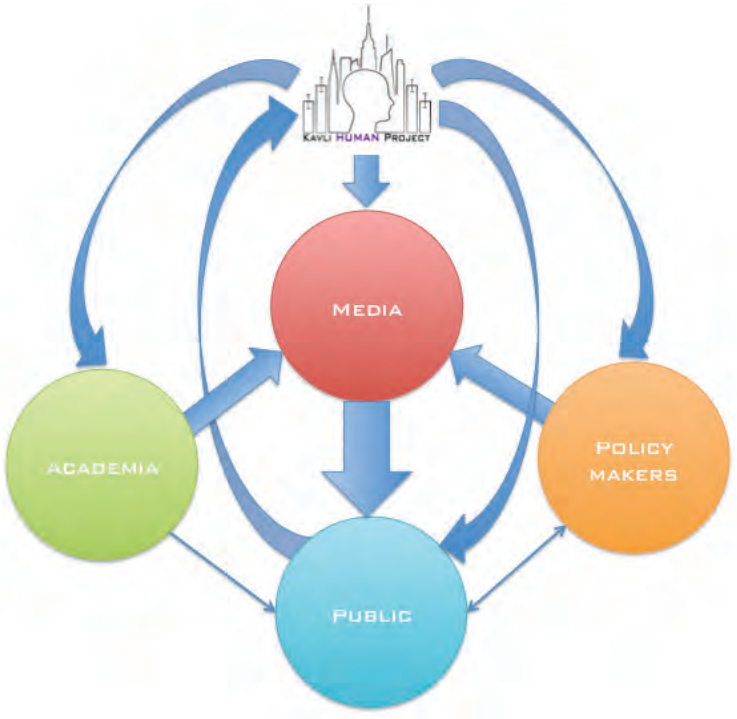


Figure 6-1: *Lines of Communication between the Kavli HUMAN Project and Key Stakeholders*

4.2 Messaging by Audience

4.2.1 Overarching KHP Message

The first and most basic messaging need will be to create simple and concise language – one or two sentences – defining both the need for and goal of the KHP. It should be a message that any audience can understand and that can be used in any context – from cocktail parties to press interviews to recruitment efforts. While it's critical to understand that public messaging is best guided by the results of quantitative message research (as discussed in the next section), a core message might incorporate some of the following language:

> *Over the last several decades, we've collectively learned how to live smarter, healthier and safer on a scientific level, but we still need to adjust our behavior, decisions, and tendencies to capitalize on that. The goal of the KHP is to bridge that gap between what we know and how we act by developing a set of best practices to help people make better decisions, while also letting them have a say in devising the policies that shape our world.*

> *Taking advantage of new technology, we will measure and analyze data types across multiple domains – genetic data, educational tracking, personality analysis, health and financial records, among many others – over a period of 20 years and then draw practical conclusions about our collective behaviors to create a set of "best practices" across an unprecedented array of situations.*

Audience-specific messages will also need to be developed. Message research (as described in the next section) will be the best way to determine the most effective narratives for each audience. Below are some potential examples of where that process could lead.

4.2.2 Academia

The academic community will be the primary recipient and end-user of the KHP database. Accordingly, initial KHP communication efforts will be geared primarily towards researchers to solicit feedback on the study to strengthen its overall design, and to foster buy-in from academia in preparation for a larger public relations campaign.

In Q3-Q4 2015 we are holding two stakeholder town hall events in Miami (September 24) and Dallas (October 23) during the annual conferences of the Society for Neuroeconomics and Economic Science Association, respectively. The Society for the Neuroeconomics—of which KHP Director, Prof. Paul Glimcher, was the founding President (2004-05) and currently serves on the Board of Directors—is the preeminent international academic society for the field of neuroeconomics. The Economic Science Association, also a major international academic organization, is dedicated to the study of economics as an observational and experimental science to learn about economic behavior.[20] The town halls, taking the form of a complementary dinner during one night of each conference, will provide a chance for KHP leadership to describe the Project and engage in a dialogue with participants to help identify strengths, weaknesses, and gauge overall receptivity of the Project. We have identified a number of different conferences to put on additional town halls in 2016 to reach other disciplines beyond economics and neuroeconomics, including the American Association of Behavioral and Social Sciences and the American Psychological Society.

In addition to these events, KHP's education and public outreach efforts will also capitalize on the work of its Scientific Agenda Advisory Council, whose members have produced multiple White Papers describing different case studies for how they plan on using the KHP dataset to advance different areas of research. These White Papers, five of which were recently published in a new issue of the journal, *Big Data*, provide ammunition for bolstering academic buy-in of the Project while allowing us to show beyond academia that: the KHP is directly involved with leading experts of their fields; that we are making serious progress; and, that there already

[20] *About the ESA*. Economic Science Association. Accessed September 14, 2015. https://www.economicscience.org/about.html

exist concrete examples for the kinds of research and societal benefits that can be derived from the KHP.

These White Papers continue to be commissioned, refined, and when possible incorporated into other journal issues. Another set of scholarly articles is expected to be published in the journal, *Economic Inquiry*, in late 2015 or early 2016. Utilizing its dedicated website, social media, outreach to different press outlets, and the leadership's personal relationships throughout multiple fields, the KHP is well poised to take advantage of these events and publications to attract and sustain additional interest in the project from across the human health and behavior research community.

Finally, the KHP website will be a resource for researchers by highlighting different grant funding opportunities from institutions like the National Institutes of Health (NIH) and National Science Foundation (NSF) to support researchers using the KHP database.

The topline messages for the academic community could be:

- *The interdisciplinary nature of the KHP offers the chance to participate in one of the most ambitious behavioral research projects ever undertaken and, subsequently, the chance to leverage the resulting data for independent research projects.*
- *The history of science shows us that whenever we create new instrumentation to study a natural phenomenon, we learn surprising new things about nature, which invariably have unforeseen, real-world applications. We don't always know* what *we will learn from something like this but we know* that *we will learn a tremendous amount.*

4.2.3 Media

The fragmentation of media presents the Kavli HUMAN Project with surmountable challenges but also with multiple opportunities to get its message out and engage with different parts of the public. When aided by the connections and clout provided by our Education and Public Outreach Advisory Council and, if possible, the support of a dedicated public relations firm, we will be well poised to capitalize on the multiple platforms and outlets available. Success in this domain requires a multipronged approach to take advantage of both traditional press platforms and newer ones—particularly social and digital media.

The Kavli HUMAN Project will need to develop relationships with many different types of media outlets, in particular local, national, and science and technology-oriented organizations. The assistance of a public relations firm will greatly help make inroads with different reports and outlets, though the KHP will not rely solely upon such a firm. Already we can leverage the many people involved in the Project for reaching out to different organizations, and have begun building relationships with places like the *Proceedings of the National Academy of Sciences*, *Scientific American*, and *Nature.* Of course, these are by and large scholarly journals and only one part of the equation, but integral to the early outreach efforts of KHP as they correlate to our efforts to reach out to the academic and research communities for feedback and buy-in in advance of a much wider public campaign.

While the science of the Project will be the primary draw for outlets like *PNAS* and *Nature*, local and national general public outlets will be more interested in the human element of the Project and the benefits for society writ large and New York City more specifically. There will also be interest in the technologies being used by the Project, which in general is a relatively more accessible topic compared to discussions of science.

Finally, social and other forms of digital media are a crucial component of the overall EPO strategy and will allow the KHP to engage with a diverse audience online, while targeted advertising to potential study populations will aid recruitment efforts. The different strategies available and resources required to accomplish our communications goals using digital and print media are detailed further in this section.

A potential top-line message for media outlets could be:

> *Doctors and scientists from many different disciplines are coming together to study humans in an entirely new way. New technologies, the big data revolution, and the maturation of multiple fields of research allow for the first time the in-depth study of human beings across multiple measurements and disciplines all at once. This big human data project of 10,000 people in New York City will create an unparalleled data platform to help us understand the roots of human behavior, how our biology affects our behavior and vice versa, and help improve the lives of [New Yorkers/the American people] through developing new theories and therapeutics, and fostering evidence-based public policies from that research.*

4.2.4 Policymakers

Long-term sustainability of the KHP will depend on leveraging resources provided by a Federal science agency, for example the National Institutes of Health (NIH). To accomplish that we will need to foster sustainable interest in the project amongst many different actors in this arena: local and state officials, Congressmen and women, the Executive Office of the President, and senior leadership at key government agencies.

Support amongst these different groups also impacts support amongst other groups. For instance, having explicit support of New York City and New York State officials and cooperation from their relevant departments will greatly enhance our ability to reach out to our potential subject pool and do so with the credibility afforded by the backing of those officials. Support from Federal actors, in turn, can help bolster support with local and state officials.

The KHP leadership has strong relationships with key players on the Federal science scene that are already solidifying the Kavli HUMAN Project's reputation in important circles. The first major meeting on the topic occurred in March 2015 when KHP leadership briefed the White House Office of Science and Technology Policy. In July 2015, Prof. Glimcher and KHP Chief Scientist, Dr. Hannah Bayer, met with directors of different Institutes of the NIH, including National Institute for Mental Health (NIMH) director, Tom Insel, and other representatives from the Executive Branch in a private meeting that epitomizes the kind of standing our leadership has. That meeting with different Institute directors led to Prof. Glimcher's invitation from NIMH director Insel to address the NIMH Advisory Council, which he did in September 2015. Prof. Glimcher was very well received by the Council members and Dr. Insel, who said that that Prof. Glimcher is "probably the smartest person I know." These initial meetings are paving the way for a sustained commitment from a Federal actor(s) to support the Kavli HUMAN Project, but complementary EPO efforts are necessary to make this happen.

Though a common trope, the government is not always on the cutting-edge. Although the research it funds often is, the government finds itself supporting such research following some sort of pull from academia and/or industry. Our outreach efforts with academia will create a groundswell behind the Kavli HUMAN Project, starting with our town hall events and leading to more scholarly articles written about the KHP, prominent academics writing op-eds in popular press, and the less publicized but equally important effect of renowned scientists reaching out directly to policymakers to talk about the KHP.

The topline messages for the policymaking/ Federal science agency community could be:

- *There is a tremendous societal benefit to the KHP: it will allow people of all social, economic, and cultural backgrounds to help themselves and their loved ones make more informed decisions about the issues that most affect them. It will similarly help inform public policy and resource allocation across the broadest possible spectrum of issues, serving as a new resource for those that work on policy proposals and legislation at every level.*
- *The particular strength of the KHP will be the extraordinary scope of the measurements it takes.*

Human behavior is an emergent property of the colliding array of internal and external factors so, to be effective, any study of it must be just as comprehensive.

4.2.5 Public

An important tenet of our public outreach strategy is an acknowledgement that the science and medical fields are not as diverse as the New York City population the KHP will study. "We are not our subjects" is the appropriate mantra. Achieving buy-in from the academic community and policymakers, combined with a strong media push, are necessary precursors to garnering public attention and support, but only a start. The Kavli HUMAN Project will use a combination of public relations tools to reach out to different neighborhoods and demographics, not all of whom have equal access to the Internet or read popular press outlets like *The New York Times*. Additionally, we will have to meet the challenge of lay publics increasingly turning to a Wild West of online sources to learn about scientific topics.[21]

In coordination with the KHP Study Frame Advisory Council, our EPO team will start the process of honing the KHP's messaging through a series of individual and group focus sessions to hear from New Yorkers what they think about the Project and what messaging they are most receptive to. With the potential for aforementioned controversy, these focus groups will allow us to understand what the average person's concerns are about the study, why they would hesitate joining the study or gladly accept an invitation to join, and provide the KHP with the opportunity to craft a successful narrative with a higher likelihood of broad appeal. The town hall format will also be adopted for the public and will allow us to reach different local communities around NYC, capitalizing on the relationships members of our Education and Public Outreach Advisory Council have with New Yorkers.

The topline message to the general public (including prospective participants) could be:

The goal of this large-scale study of human behavior is to shine a spotlight on the biggest problems plaguing communities across New York City—how they impact our daily decisions and how to make them better. KHP researchers have developed a way to let everyday New Yorkers improve their own lives while also playing a key role in shaping the policies that will guide our city.

4.3 Reaching Target Audiences: Events and Timing

Different outlets reach different audiences. In order to drive the right narrative with academia and agencies/policymakers, earned (and paid) media in scientific journals and political trades are the primary focus; hyperlocal and NYC outlets are among the most visible options to complement recruitment efforts of study participants; national media hits will shape the overarching public perception about the KHP. We will also explore special events like a science communication workshop with Alan Alda and his Center for Communicating Science. Below are both broad and specific suggestions for the best way to target each of the audiences identified above.

[21] Scheufele, D.A. (2014). Science communication as political communication. *Proceedings of the National Academies of Sciences,* Vol. 4, Suppl. 4. http://www.pnas.org/content/111/Supplement_4/13585.full.pdf

Table 6-1: *Target Audience Resources*

Academia	**Federal Scientific Agencies / Policymakers**	**Prospective Participants/ All Audiences**
• Scientific Journals o JAMA o NEJM • Science Media Outlets o Science Magazine o Popular Science o National Geographic o Discovery Channel o NPR • TED Talks	• Political Trades o The Hill o Roll Call • Policy Journals o Health Affairs o Journal of Health Economics • Local Press o hometown paper of Chair of Labor, Health and Human Services Cmte., which oversees NIH funding (currently OK Rep. Tom Cole) o hometown paper of Kavli Foundation board members • Digital Media (See Section VII)	• Hyperlocal Media o Staten Isl. Advance o Queens Tribune o Bronx Beat o News 12 Bronx/Bk. • Mainstream NYC Media o Daily News o WNYC o NY1 • National Media o NYT o WSJ o CNN o CBS/NBC/ABC News

Sequencing the audience-specific messaging will be a critical component of the overall communications plan. Below is a notional timeline:

Table 6-2: *Notional Public Outreach Campaign Timeline*

	2015	**2016**	**2017**
1st Quarter	–	• Begin IDIs • Begin Polling • Legislative outreach o Lobbyist?	• Recruit 10 test families • Hyperlocal media hits appear as national hits continue • Paid media
2nd Quarter	–	• Focus groups • Science communications event with Stony Brook Alan Alda Center for Communicating Science • World Science Festival public unveiling o NYT or other major press outlet profile	• Door-knocks begin
3rd Quarter	• Finalize communications plan • Town Hall at Society for Neuroeconomics Fall 2015 Conference • Town Hall at Economic Science Association 2015 Fall Conference	• Begin outreach to mainstream media hits • Speaking engagements and interviews with KHP senior leadership	*(TBD)*
4th Quarter	• Finalize message research plan • Academic outreach o Science journals • Paid Media	• Mainstream media hits appear • Begin outreach to hyperlocal reporters • Local Community Town Halls	*(TBD)*

5. Researching Outreach Strategies and Messaging

Message research and testing will validate – and drive – KHP's study efforts. KHP has a tremendous challenge in not only recruiting a very socio-economically, gender, and age diverse pool of candidates, but also in convincing participants to provide the most intimate access to the details of their or their children's lives. The KHP isn't just trying to win hearts and minds on an issue; it must convince families and individuals to allow unparalleled access to their bodies, finances, and most private moments.

This section is defined as message research but, in actuality, the research will also need to encompass measuring the resonance of the KHP brand as well as the brand of the study. People might be less willing to participate in something under a brand they have just come to learn about versus a brand or message they have long experience with. This is especially true with the level of intimacy KHP expects from study participants. Additionally, this research will need to explain how the individuals in the different demographics make trust and value decisions; simply stated, the research will need to uncover who or what validates their decision making and where in their community do they go for this type of information.

It is impossible with a finite budget to test and examine the motivations of every category of New Yorker at the depth needed to fully realize a personalized complex value proposition for each participant. However, the study frame will provide us with details on the study population by language and neighborhood, allowing the KHP to efficiently hyper focus on specific cultural and linguistic realities, making the task of researching participant motivations more doable. Additionally, when used correctly, research can also be used not only as a means to gather data on a value proposition, but also as a real-time and evolving means of educating gatekeepers (i.e. community leaders) as to KHP's importance, ethical and privacy standards, etc., helping to build recognition of KHP in the communities in which participants are needed.

The research recommendations laid out below outline how the KHP's education and public outreach efforts can best complement study participant recruitment and ensure the widest possible population is reached, including vulnerable, less digitally inclined sub-populations, as well as the elderly and children (via their parents). This approach allows the KHP to test and refine in real-time, as well as begin the education process with thought leaders who have constituencies that we want to reach. Note that all of this is very scalable; if the budget prescribed is a non-starter, research scope can always be recalibrated. However, with the weight of the ask for participants and the diversity of the participant sought, it is important to fund as much research as possible—beginning in 2016—without shortchanging other elements.

Step One: Influencer In-Depth Interviews (IDIs)

The first step will be to identify 20-30 key influencers and members of the community across a broad spectrum of socio-economic categories that fall into two categories (with some overlap between both of them): 1) those that understand the need for the type of dynamic data the KHP project can generate and can see how people could be helped by it and 2) those that are part of the communities where we believe recruitment might be more of a challenge. For example, our 20-30 IDIs might fall across the following categories:

- 4-6 New York City residents who work in health care (including some who specifically work with the elderly).
- 4-6 New York City residents who work in education (including some who specifically work with children).
- 4-6 New York City residents who work for or volunteer with community non-profit organizations.

- 4-6 New York City residents who work or volunteer with immigrant populations in New York City.
- 4-6 New York City residents who say their family/friends/peers looks to them for advice on important matters.

Some of the topics to explore include:

- How should the project present itself to participants? (Test "the Kavli Human Project" against other possibilities like "the Human Project," etc.)
- When presented with the information, what are people's initial impressions of the project and its overall goals?
- What are the biggest motivators to participate?
- Sympathy for the goals of the project.
- Attitudes towards financial compensation.
- The impact of celebrity participation/exclusivity of the sample.
- Tapping into issues that matter for sub-populations.
- What are the biggest barriers to participate? Health concerns? Financial concerns? Privacy concerns? Parental permission for children? Elderly participation?
- Identity and role of key influencer channels.
- Role of social media/advertising channels.

These interviews will establish a baseline understanding of the issues involved in recruiting for this project and will help to surface both anticipated and unanticipated barriers to participation. IDIs will help us understand what brand identity or messages will establish a baseline of trust for individuals to even consider participation, and the voices and opinions that these individuals will seek as resources in their decision-making process. It will also assist us in formulating the language that will be the most effective in convincing residents to participate and creating opportunity bridges to recruit their friends and family. Lastly, it will help us begin conversations with gatekeepers, build awareness of the KHP project, and create de facto spokespeople that can speak at length about the importance of this project.

Step Two: Online Quantitative Survey

The IDIs will be followed by an online quantitative survey of New York City residents that will allow for analysis of results by demographics, psychographics, behavioral characteristics, as well as by borough. This survey would seek to test and confirm hypotheses developed during the initial IDIs and preliminary messaging based on those hypotheses. Initial drafts of messaging and creative materials could be developed that would be tested at the end of the online survey instrument. This messaging would address which underpinning motivations will be most effective at driving participation and engagement, as well as how to counter barriers such as privacy concerns, etc. For example, which prospective participants are likely to be motivated by the potential for self-improvement? By an altruistic sense of bettering society? By a "cool factor" of being involved in a cutting-edge project? By the financial compensation?

Step Three: Focus Groups

Lastly, once the survey is completed, further testing of more evolved messaging and creative work would be undertaken to prepare for the final launch of the recruitment phase. Ideally we will conduct 8-10 focus groups with a diverse range of audiences as a final check on the messaging strategy and execution of the recruitment messaging and materials. Focus groups are so essential because KHP (via the moderator) can have a deep two-way discussion that generates more open-ended information on efficacy of the messages, rather than forcing people into standardized responses. Focus groups allow for real-time tinkering of questions if unexpected information arises in the discussion via firsthand observation and can be a great way to uncover more sophisticated and nuanced reactions to the almost finalized messaging. Audiences could be grouped by:

- Gender
- Age
- Borough
- Ethnicity
- Language spoken

6. Paid Media Strategy

Paid communication can build visibility among key stakeholders and support participant recruitment both by "priming" New Yorkers with credible, positive information about the KHP and by persuading undecided participation targets after an initial visit by the enrollment team.

The following specific recommendations are driven by a series of principles, predicates and assumptions—about both the recruitment plan and the use of paid media in New York City:

- The extraordinary cost of media in the city places a high premium on efficiency, meaning use of highly-targeted mediums.
- The enrollment team will be traveling through the city, geographic cluster by geographic cluster, targeted down to the building or individual housing unit level. This lends itself well to paid communication tools that can be similarly targeted.
- Recruitment will occur over a rolling 36-month period (or longer). The "rolling" approach to recruitment lends itself well to highly-targeted communication that "rolls" with the enrollment team.
- Generating sustained recall of a brand is both costly and difficult. To the extent possible, recruitment targets should see communications in the 1-3 week period immediately preceding the arrival of the enrollment team at their home.
- The enrollment team is likely to be re-contacting "undecideds." Paid communication can follow these New Yorkers after the first contact (but before the second contact) with an appropriate persuasion message.

Choice of Media:

Mass media—broadcast and cable TV, terrestrial and satellite radio and citywide newspapers—are powerful engines for branding and persuasion. However they are prohibitively expensive to produce recall that is sustained for 12 months or longer. They are also highly inefficient—paying to reach almost 20 million people in the tri-state area when the KHP may be targeting as few as 100,000 households.

Instead, it would be more efficient to devise a media strategy that targets, for example, the 1,000 households in the Crown Heights neighborhood whose doors may be scheduled to be knocked on in the first week of April 2017, followed by 1,000 households in Bedford-Stuyvesant in the second week, followed by 1,000 households in Clinton Hill in the third week. In other words, design a highly-targeted paid media strategy to reach people before/just as the recruitment team approaches.

It is also possible to communicate directly with participant targets that, upon their first visit, request more time to consider whether or not to participate in the study.

Specific Suggestions:

- **Digital Advertising (both video and display)**
  This can be targeted down to the level of an individual person with a relatively high rate of accuracy (using information that many of us provide to digital advertising networks to allow super-precise "cookies" to be placed on our computers). A smart strategy would mix individual-targeted ads with zip code-targeted (or even more narrowly geo-targeted) ads designed to reach the portion of the target population who are not individually targetable. Advertising would begin about two weeks before the enrollment team is scheduled to arrive at a target's home.
- **Direct Mail**
  Mailings can be targeted on an individual or household level and mailed on a rolling basis in advance of the enrollment team's arrival across (or within) a specific community. Mail has the potential to reach those who are truly unreachable through any digital channel.

- **Print Advertising in Local/Ethnic Newspapers**
  New York City's diverse communities have a wide array of local, ethnic, and foreign language newspapers – many of which are well read in certain neighborhoods. They are relatively costly and would not be as targeted as digital and direct mail approaches, but could be of value in priming specific audiences in a higher-end budget.

No method is perfect. The principal drawbacks of digital advertising are its limited reach with older participation targets, challenges matching language to audience, and the evolving threat of ad-blocking and consumer-driven obstacles to super-precise targeting. Direct mail can easily be ignored as "junk mail" and is less likely to be read by younger New Yorkers. Local print advertising is very costly on a "CPM" basis and has limited reach. A multi-medium strategy as laid out above will yield the best results.

7. Crisis Management and Rapid Response

Given its tremendous scale and the inherent concerns about the privacy of research subjects, the Kavli Human Project must take steps to prepare both for tough questions about its privacy safeguards, as well as the possibility of a larger communications crisis, such as a partial or full data breach.

As discussed in Section III above, digital privacy and security are a top concern and favorite topic of local and national news media. Once a story gathers momentum, even the facts are not always enough to slow it down. The following proactive and reactive strategies will help overcome these challenges and prepare for the pitfalls inherent to any project of this scale.

Proactive Communication

Many problems can be avoided with clear, transparent communication about the KHP's work. Study leaders must proactively communicate – both with research participants and the media – about how the data will be collected, stored, used, and disseminated and the rigorous steps that the Project is taking to ensure that the privacy of its participants and their data is maintained throughout. These policies should be listed (1) on KHP's website in clear, easy to understand FAQ format, in multiple languages as discussed above and (2) distilled into a one-pager that is proactively shared with participants and can be used as background material with media.

Reactive Communication

As a supplement to the communications materials above, the KHP must prepare detailed, rapid response fact-sheets and Q&A materials that clearly explain its data security procedures and privacy policies in full technical detail. While such information may ultimately never be needed, the Project must be prepared to fully explain its practices to the media, and potentially to authorities and/or third-party data security and privacy experts, should the need arise.

Crisis Communication / Data Breach

To prepare for a communications crisis like the leak of personal data, the KHP must develop and implement a detailed response protocol, informed by the mistakes and successes of similar incidents (like those discussed in Section III).

- *Communications Principles:* In any crisis scenario, the Project must ensure that its communications with participants and the press are clear, accurate, and transparent. Many mistakes can be avoided simply by being straightforward and upfront about what is known and what is not.
- *Procedures:* In addition to the procedures developed by the data team to collect and store participant data, the Project must install protocols to monitor for vulnerabilities and breaches, and determine ahead of time what errors may require communication with participants and the public.

- *Staffing and resources:* Successfully managing a crisis means being able to move and react in real time. The Project's data and communications teams must be sufficiently resourced and staffed so that they have the capacity to identify problems as they occur, communicate facts quickly and clearly, answer questions, and implement solutions in a timely manner. During a communications crisis, the Project team should be able to devote significant staff time to addressing the issue.
- *"War Room:"* Moving past a communications crisis requires close coordination and campaign-style tactics to correct the record and respond to media criticism. KHP staff should be ready, if needed, to execute a daily War Room conference call, including:
    - Project leadership;
    - Communications team;
    - Technical / data collection team;
    - Responding in real time to media reports, reporter inquiries;
    - Planning corrective action if and where errors have been made.

8. Governance: Education and Public Outreach Advisory Council

We are also in the process of standing up the EPOAC to support this area and complement our other advisory councils in Measurement Technology, Privacy & Security, and Study Frame. The EPOAC seeks to educate key constituencies, including academia, the press, the public, and policymakers on research and findings of the Kavli HUMAN Project. The EPOAC will oversee implementation of KHP's communications strategy, provide access to media outlets and thought leaders, and attract and retain the interest of sponsors. In addition, the EPOAC will assist in recruiting members to the Board of Directors to provide visible public support for the project in the popular press. Members will also act as conduits to provide feedback to KHP leadership from the many audiences they are in touch with.

Membership of the EPOAC will necessarily be well rounded to be able to talk about the many scientific, technical, and public policy aspects of the study with our diverse set of stakeholders. Council members will be experts in medicine, science, publishing, public relations, and more specifically science communication, who will bring clout among a wide range of audiences to advocate for the KHP and the positive change it can foster; help create and fortify relationships with stakeholders; and fight back against negative coverage.

Some Council members will come from the social justice and activist communities to help represent the perspectives of underserved minorities, as the study sample will be as representative of the whole population of New York City as possible. Their work in that area will help us reach those communities who often have less access to digital media and the Internet and are harder to reach out to.

While we have already lined up a number of council members from the academic and science communication arenas, unlike the other KHP advisory councils, there is a special onus on the EPOAC to recruit as its chair someone of public (not just academic) renown who can help with additional council recruitment, among other duties. Given that we are creating a brand from scratch that will require significant public participation, having a publicly-recognizable figure on the EPOAC can enhance the public's receptivity to the KHP as well as the success of its messaging. Recruiting a high-level public figure will require a concerted effort between the KHP team, the Kavli Foundation, and New York University to find someone of sufficient stature for the role of council chair.

9. Budget

10. Conclusion

Other sections of this Preliminary Design document demonstrate the import of their domain to the success of the Kavli HUMAN Project, and education and public outreach is no exception. As one of the five pillars of the KHP, the success of the Project will depend on positive reception from academia, government, the public, and the media. The Kavli HUMAN Project has given considerable thought to how it should engage with a wide array of stakeholders and audiences and explain why the KHP is such a compelling and revolutionary effort. Not only is it important to help people understand the purpose and goals of the Project, but also to convey the excitement of those involved in the Project. At the same time, the Project team has considered the challenges that could potentially crop up and will be able to mitigate certain risks while being ready to address other issues as they happen.

Accomplishing all of this requires carrying out the strategy described in this section: message research; holding different types of outreach events with diverse audiences; paid media; and standing up an advisory council that represents different disciplines, industries, groups, and neighborhoods across New York City.

The aspiration and promise of the KHP is a powerful tool in itself, but will be combined with rigorous preparation to handle all manner of situation. The Kavli HUMAN Project is eager to begin an open dialogue with the public and looks forward to the opportunity to learn from its many stakeholders.

7

SCIENTIFIC AGENDA

1. Introduction

In order to represent the breadth of use that the Kavli HUMAN Project data will have, and the need for this kind of data collection, experts from diverse academic specialties have written to reveal specific research uses for KHP data. 18 papers have been commissioned and confirmed to this end. Current papers include work in economics, psychology, biology, neuroscience, and law, with tangential connections to various other fields. While these papers are in an array of stages of completion, 7 subgroups have been identified for a tiered publishing plan. Five papers have been published as a set in Big Data. Another article has been submitted for review in Economic Inquiry, to be followed by 4 other papers to make a set. A third set of psychologically-focused papers have been commissioned to appear as a set of 4. Four individual articles have been commissioned, with intent to publish singularly in law, neuroscience, biology, and environmental change. The 5 published papers appear in their published format in Appendix K, and in-progress drafts of the other papers follow that set. For recently commissioned papers that are not yet fully developed, working titles are provided below.

2. Summary Report of White Paper Status

(As of October 1, 2015)

2.1 Published (5)

The Kavli HUMAN Project: Using Big Data to Understand the Human Condition

Citation: *Big Data*. September 2015, 3(3): 173-188; DOI: 10.1089/big.2015.0012

Authors:

Okan Azmak, Measurement Technology Officer, Kavli HUMAN Project;

Hannah Bayer, Chief Scientist, Institute for the Interdisciplinary Study of Decision Making, Research Associate Professor of Decision Sciences, New York University;

Andrew Caplin, Deputy Director, Institute for the Interdisciplinary Study of Decision Making, Silver Professor of Economics, Department of Economics, New York University;

Miyoung Chun, Executive Vice President of Science Programs, The Kavli Foundation;

Paul Glimcher, Director, Institute for the Interdisciplinary Study of Decision Making, Julius Silver Professor of Neural Science, Economics and Psychology, New York University;

Steven Koonin, Associate Director, Institute for the Interdisciplinary Study of Decision Making, Director, Center for Urban Science + Progress, New York University;

Aristides Patrinos, Member, Board of Directors, Kavli HUMAN Project

Key Points:

- There is a need for a single repository containing something like a complete record of the health, education, genetics, environmental, and lifestyle profiles of a large group of individuals at the within-subject level, and this repository may now be possible.
- The Kavli HUMAN Project (KHP), an effort to aggregate data from 2,500 New York City households in all five boroughs (roughly 10,000 individuals) whose biology and behavior will be measured using an unprecedented array of modalities over 20 years, may be the answer to this need.
- The KHP will provide synoptic and granular views of how human health and behavior co-evolve over the life cycle and why they evolve differently for different people. This will enable new discovery-based scientific approaches, rooted in Big Data analytics, to improve the health and quality of human life, particularly in urban contexts.

Real Time Assessment of Wellness and Disease in Daily Life

Citation: *Big Data*. September 2015, 3(3): 203-208; DOI: 10.1089/big.2015.0016

Authors:

Dennis Ausiello, Jackson Distinguished Prof. of Clinical Medicine, Director of Harvard Medical School's M.D./Ph.D. Program, and Emeritus Physician-in-Chief of Harvard Medical School. Member, IOM & AAAS;

Scott Lipnick, Scientific Director, Center for Assessment Technology and Continuous Health (CATCH); Assistant in Biomedical Physics, Department of Medicine, Massachusetts General Hospital, Imaging and Data Specialist, Stem Cell and Regenerative Biology Department, Harvard University

Key Points:

- The next frontier in medicine involves better quantifying human traits, or "phenotypes".
- Biological markers have been directly associated with disease risks, but poor measurement of behaviors such as diet and exercise limit our understanding of preventive measures.
- By joining together an uncommonly wide range of disciplines and expertise, KHP will advance measurement of behavioral phenotypes, as well as environmental factors that impact behavior.
- By following the same people over time, KHP will liberate new understanding of dynamic links between behavioral phenotypes, disease, and the broader environment.
- As KHP advances understanding of the biology-behavior nexus, it will seed new approaches to the diagnosis, prevention, and treatment of human disease.

Life-Course Risks and Outcomes of Cognitive Decline

Citation: *Big Data*. September 2015, 3(3): 189-192; DOI: 10.1089/big.2015.0015

Authors:

Kenneth Langa, Prof. of Medicine, UMich; Research Scientist, UMich Veterans Affairs HSR&D Center for Clinical Management Research; Assoc. Director, Institute of Gerontology. Member, American Society for Clinical Investigation; Member, Health and Retirement Survey.

David Cutler, Otto Eckstein Professor of Applied Economics, Harvard; Research Associate, NBER; Council of Economic Advisers and National Economic Council (Clinton); Presidential Campaign Advisor (Bill Bradley, John Kerry, Barack Obama); Senior Healthcare Advisor, Obama Presidential Campaign.

Key Points:

- In 2010, about 4.2 million adults in the U.S. had dementia with an economic impact estimated at $200 billion per year, with both numbers projected to increase rapidly.
- The KHP will provide uniquely rich measurements of cognitive decline, as well as how caregiving impacts caregivers and health care utilization.

- By adopting a full life-course approach, KHP will measure risk factors for later-in-life cognitive decline, including earlier cognitive activity, health, social interactions, and work status.

How Genetic and Other Biological Factors Interact with Smoking Decisions

Citation: *Big Data*. September 2015, 3(3): 198-202; DOI: 10.1089/big.2015.0013

Authors:

Laura Bierut, Alumni Endowed Prof. of Psychiatry, Co-Director, Outpatient Clinic, Washington Univ. in St. Louis School of Medicine. Member, NIDA Genetics Consortium; Lead, Collaborative Genetic Study of Nicotine Dependence.

David Cesarini, Asst. Prof. of Economics, Center for Experimental Social Science, NYU; Co-Director, Social Science Genetic Association Consortium.

Key Points:

- Despite clear links between genes and smoking, effective public policy requires far richer measurements of the feedback between biological, behavioral, and environmental factors.
- By convening a wider range of expertise than is traditional, KHP will allow researchers to paint a much richer picture of an individual's lifecycle trajectory of smoking (and other substance abuse) and interactions with other choices and environmental factors.
- The longitudinal nature of KHP will be particularly valuable in elucidating these insights, particularly in light of the increasing evidence for how smoking behavior affects physiology and health.

Diet, Economics, and Health in the Family and Community Context

Citation: *Big Data*. September 2015, 3(3): 193-197; DOI: 10.1089/big.2015.0014

Authors:

Adam Drewnowski, Prof. of Epidemiology; Director: Nutritional Sciences Program, Center for Public Health Nutrition, and Center for Obesity Research, Univ. of Washington; Public Trustee, International Life Sciences Institute; Inventor, Nutrient Rich Foods Index & Affordable Nutrition Index.

Ichiro Kawachi, John L. Loeb & Frances Lehman Loeb Prof. of Social Epidemiology, Chair of Dept. of Social & Behavioral Sciences, Harvard; Co-Director, Robert Wood Johnson Foundational Health & Society Scholars; Chair, Harvard School of Public Health's Institutional Review Board; Member, IOM.

Key Points:

- Teasing apart the links between neurobiology, economics, culture, and the food environment is of immense importance to public health. Yet answers to even such a basic question as whether or not a healthy diet necessarily costs more remain unclear.
- The KHP will advance our knowledge of links between food choice, economic status, and biological, environmental, and social factors. It will create a first-of-its-kind "nutrition atlas" to understand New Yorker's current food decisions, and to identify the changes that would most improve the quality of their diets.
- The KHP will advance our knowledge of how nutritional awareness and chronic conditions (diabetes, obesity) impact food shopping behaviors.

2.2 In Review (1)

Family Decision Processes and the Quality of Secondary Schooling Decisions

In review at: *Economic Inquiry*

Authors:

Pamela Giustinelli, Research Assistant Prof., Institute for Social Research (Survey Research Center), UMich; Affiliate, Michigan Center on Demography of Aging and Michigan Institute for Data Science; Member, Human Capital & Economic Opportunity Global Working Group, Univ. of Chicago; Member, American Economic Association.

Charles Manski, Board of Trustees Prof. of Economics, Northwestern Univ.; Member, NAS; Fellow, Econometric Society, AAAS, and British Academy.

Key Points:

- KHP will fill glaring gaps in our knowledge of how families and students make their increasingly complex secondary education choices.
- This research will use highly successful survey methodologies for measuring expectations employed for over two decades by major household studies like the NIH-funded Health and Retirement Study.
- Research findings will enable policymakers to improve New Yorkers' understanding of school options and more effectively allocate educational resources.

2.3 In Progress (12)

Long-Term Care in the Family Context

Proposed Venue: *Economic Inquiry*

Authors:

Andrew Caplin, Silver Prof. of Economics, NYU; Chair, Kavli HUMAN Project Scientific Agenda Advisory Council; Dep. Director, NYU IISDM; Research Associate, NBER; Fellow, Econometric Society.

Kathleen McGarry, Chair, Dept. of Economics, UCLA; Research Associate, NBER; White House Council of Economic Advisers (Clinton).

Key Points:

- Long-term care bears with it great financial uncertainty and stress, placing a burden on older Americans and their families engaged in such support.
- The KHP will provide the first holistic picture of how long-term care impacts the individuals involved and their families.
- It will illuminate differences between outcomes for individuals without a support structure and those with familial support.
- The resulting knowledge will not only improve public policies, but also the design of insurance options available in the private market place.

Measurement and Urban Spaces

Proposed Venue: *Economic Inquiry*

Author:

Edward Glaeser, Fred & Eleanor Glimp Prof. of Economics, Harvard; Director, Rappaport Institute for Greater Boston; Senior Fellow, Manhattan Institute; Fellow, World Bank; Member, Gates Foundation Program Advisory Panel; Research Fellow, NBER; Member, AAAS.

Key Points:

- Do cities encourage the free flow of information? How does the geography of food availability influence diet? Answers to these and other questions in urban economics require novel data gathering that has geography and community at is center.
- With its highly detailed measurement of the life of New York City and its residents, the KHP will answer key questions of how geography interacts with behavior and knowledge.
- In addition to tracking individual behavior over time, the KHP will study changes in population, resident flows, buildings, amenities, and

property values. This panel aspect will prove particularly valuable given the rich changes—anticipated and unanticipated—that the city will undergo in the next generation.

Integrated Household Financial Surveys *(Working Title)*

Proposed Venue: *Economic Inquiry*

Authors:

Robert Townsend, Elizabeth and James Killian Professor of Economics, Department of Economics, Massachusetts Institute of Technology (MIT)
Scott Schuh, Director, Consumer Payments Research Center, Senior Economist and Policy; Advisor, Research Department, Federal Reserve Bank of Boston

How to Measure Spending *(Working Title)*

Proposed Venue: *Economic Inquiry*

Authors:

Robert Townsend, Elizabeth and James Killian Professor of Economics, Department of Economics, Massachusetts Institute of Technology (MIT)
Scott Schuh, Director, Consumer Payments Research Center, Senior Economist and Policy; Advisor, Research Department, Federal Reserve Bank of Boston

Life Adversity during Sensitive Periods of Brain Development: Social & Economic Impact

Proposed Venue: *Developmental Science*

Author:

Regina Sullivan, Prof. of Child and Adolescent Psychiatry, New York University School of Medicine; Research Scientist at Nathan S. Kline Institute for Psychiatric Research.

Key Points:

- Repeated trauma and abnormally high, prolonged stressful experiences, particularly adversity from caregivers, target the development of emotional and cognitive circuits in the brain. Environmental neurotoxins like pollution can amplify these effects.
- Recent work suggests that social support, particularly from caregivers, can buffer the effects of adversity on the developing child's brain.
- Capitalizing on new technology, the KHP will assess the amount and quality of social contact between the child and others in the home environment, suggesting potential mechanisms by which the social interactions can influence childhood brain development.

The Role of Family, Neighborhood and Educational Context in Child Development *(Working Title)*

Proposed Venue: *Developmental Science*

Author:

Jeanne Brooks-Gunn, Virginia and Leonard Marx Professor of Child Development and Education, Teachers College and College of Physicians and Surgeons, Columbia University, Co-director, National Center for Children and Families, Co-director, Columbia University Institute for Child and Family Policy

Child Development *(Working Title)*

Proposed Venue: *Developmental Science*

Authors:

Clancy Blair, Professor of Cognitive Psychology, Department of Applied Psychology, Steinhardt School of Culture, Education, and Human Development, New York University
C. Cybele Raver, Vice Provost for Academic, Research, and Faculty Affairs, New York University

Psychological Development in Adolescents
(Working Title)

Proposed Venue: *Developmental Science*

Authors:

BJ Casey, Director, Sackler Institute for Developmental Psychobiology; Professor of Developmental Psychobiology, Weill Medical College of Cornell University
Catherine Hartley, Assistant Professor of Psychology in Psychiatry, Sackler Institute for Developmental Psychobiology, Weill Medical College of Cornell University

When Is It Important To Make Choices?

Proposed Venue: Law Publication

Author:

Cass Sunstein, Robert Walmsley University Prof., Harvard; Director, Program on Behavioral Economics and Public Policy, Harvard Law School; Member, Bloomberg Government Advisory Board; Former Administrator, White House Office of Information and Regulatory Affairs (Obama); Former Law Clerk for Supreme Court Justice Thurgood Marshall

Key Points:

- Many people avoid actively choosing, instead engaging in "default" behaviors. Businesses and governments are increasingly aware of this inertia, yet the research community has done little to measure the impact on a larger scale in field settings.
- KHP offers the chance to study the interaction between default rules and behavior in the profoundly dynamic urban context of New York City. This will improve our understanding of how messages of various kinds can produce positive behavioral change, and how to counter messages that promote potentially damaging behaviors.
- The dynamic aspect of KHP will enable researchers to gauge the extent to which active choosing promotes learning and thus the development of preferences and values.

Longitudinal Studies in Neuroscience – Understanding Human Dynamics

Proposed Venue: *Nature Neuroscience*

Author:

Russell Poldrack, Prof. of Neuroscience and Cognitive Psychology, Stanford University;

Key Points:

- Psychology and neuroscience currently have almost zero knowledge about how the mind and brain change over time scales of days, weeks and months.
- This is particularly problematic given that most major mental health disorders (including depression, schizophrenia, and bipolar disorder) involve drastic fluctuations in mental function over the course of weeks.
- KHP will allow researchers to examine changes in behavior and mental function over time scales of days, weeks, months and even years. Such data will provide the opportunity to link these relatively rapid changes to mental health disorders, as well as to provide new insights into existing research on slow changes previously identified in studies of child development and aging.

Rewards and Economic Decisions: From Lab to Field

Proposed Venue: *PLoS Biology*

Author:

Wolfram Schultz, Wellcome Principal Research Fellow, Prof. of Neuroscience, Univ. of Cambridge; Fellow, Royal Society; Recipient, Ellermann Prize, Theodore-Ott Prize, Golden Brain Award, Ipsen Prize.

Key Points:

- Neuroscientific research has traditionally been limited to laboratory settings, precluding many kinds of research that could help advance, and even revolutionize, the field.
- KHP will allow researchers to determine real-world applications of laboratory work, in particular work on the dopamine system and reward prediction errors, and our mechanisms for self-control.
- KHP will help identify circumstances under which reward signals overcome self-control mechanisms, as may occur in such cases as drug addiction and in the dietary arena.
- KHP will enrich our understanding of how environmental conditions and differences across individuals impact the balance between the pull of reward and opposing self-control mechanisms.

Bridging Local & Remote Earth Sensing Platforms to Understand Feedback Mechanisms Between Local Populations and Environment ***(Working Title)***

Authors:

Aristides Patrinos, Member, Board of Directors, Kavli HUMAN Project; Fmr. Deputy Director for Research, Center for Urban Science + Progress, New York University

Masoud Ghandehari, Head, Urban Observatory, Center for Urban Science + Progress; Associate Professor, Civil & Environmental Engineering, Polytechnic School of Engineering, New York University

2.4 Recently Invited (1)

Energy Efficiency *(Topic)*

Author:

Steven Koonin, Director, Center for Urban Science + Progress; Associate Director, Institute for the Interdisciplinary Study of Decision Making

APPENDICES

APPENDIX C

ADVISORY COUNCIL REPORTS

The following reports are historical accounts of three Advisory Council meetings with leading experts in the fields of Measurement and Technology, Study Frame Design, and Privacy and Security; held on November 24, 2014, December 12, 2014, and January 20, 2015 respectively.

APPENDIX C - 1

MEASUREMENT AND TECHNOLOGY ADVISORY COUNCIL WORKSHOP SUMMARY REPORT

November 24, 2014

Executive Summary

The Measurement and Technology Advisory Council met on November 24th, 2014 to discuss the measurement design for the HUMAN project. We discussed the list of desired measurements as well as some of the technical issues surrounding data collection. The highlights of the resulting design are described in the text below and a detailed outline that follows.

List of Measurements: A preliminary list of desired measurements is now complete and follows in detail at the end of this document. Measurements can be divided into: 1) Capture of physical samples, 2) Psychological assessment, 3) Social network and communication pattern profiling, 4) Location data, 5) Health data, 6) Education data, 7) Employment data, 8) Financial data, and 9) Socio-political assessment (a broad category encompassing voting records, religious and philanthropic activity).

Importance of Developing a "Time-Budget": An important consideration going forward will be the week-by-week budget for use of participant's time. The budget will include a pre-enrollment process, about a month prior to official intake, during which we will learn about participants' current technology use and begin some preliminary psychological profiling via questionnaires. The enrollment process is now slated to take one day for physical sample collection and other initial intake processes. We will also develop a weekly, monthly, and annual budget time for surveys delivered to participants' smartphones in game-ified form.

New Measurement Concepts: Several new measurement concepts emerged that will be folded into the current study design documents. These are: 1) Automated life or behavioral event triggered data gathering, 2) Providing a mechanism for additional investigator-initiated measurements of panel subgroups, 3) Fitting of Bluetooth beacon bracelets to children < 10yrs of age, 4) Marking household locations with Bluetooth beacons, and 5) The need for initial focus group analysis of measurement acceptability prior to the CD1 document completion.

Study Frame: Although the detailed composition of the participant pool remains to be specified by the Study Frame Design Council, the measurement group argued for a three-tiered participant pool:
1) Comprehensive measurements on 3000 families, 2) More detailed measurements (e.g. brain scans) in a group of 300 people, and 3) A larger citizen-scientist population of ~300,000 participating through the internet.

With these basic design elements now in hand, we can turn to cost-benefit analysis and prioritization of the potential measurements, with the goal of identifying the optimal set of measurements for each subject pool.

Detailed Summary

Key Discussion Points

The Time Budget for Subjects

Although we expect that the bulk of the data will be collected automatically following the enrollment process, there will still be some forms of data collection for which the subjects will have to be active participants (e.g. social network self-reports or surveys about major life events). We propose to develop a time budget for participants on a weekly, monthly, and annual basis that balances our data collection needs with the amount of time that participants can reasonably be expected to give – and be compensated for. This budget will include a range of time demands from shorter, more frequently gathered items like surveys, to longer, more invasive procedures like updating biological samples. Frequent events might be gathered by internet on a weekly basis, while more detailed procedures gathered as rarely as annually or biennially.

The enrollment process will be a substantial physical undertaking. In a setting like a storefront medical clinic, we expect that a fully trained intake staff could enroll approximately 5-10 families per day. Even with two of these enrollment clinics working at full capacity, given a reasonable rate of attrition, we expect that it will take 3 years to completely enroll the study. While these clinics will be sited to optimize accessibility to residents throughout New York City, we may be able to improve our reach, particularly for those like the elderly and low socioeconomic status individuals, by going to them with a mobile medical vehicle. One of these mobile clinic-like vehicles could be placed each week in a different neighborhood. This would allow us access to 50 neighborhoods per year – each 'neighborhood' covering about 3 zip code tracts. Such an approach could be highly cost-effective.

New Measurement Concepts

In addition to these regularly scheduled data collections, we also envision performing life event-triggered sampling. In the most straightforward case, we can follow up with participants any time that they report an event on one of the regular traumatic life event surveys. However, we could also trigger data collection protocols on large changes in the incoming data streams, such as radical changes in the nature of a subject's social network or physical location. The development of sample triggers is proposed as a major focus of the CD2 process.

Technology for data collection will undoubtedly evolve rapidly over the course of the study. Therefore, it is important that we specify the kind of data that we would like to gather rather than the technology that we use to collect it. For example, we hope to gather information about subject location by the most accurate method available, whether this is via GPS, wireless, or Bluetooth signaling. This kind of conceptualization will allow us use of the optimal technology available to meet our needs, despite the changes that will surely occur as the study progresses. One technology that emerged as worthy of detailed consideration is Bluetooth beacons. These small devices, with battery lives of up to 3 years, could provide powerful tools for tracing children and for highly accurate geo-location.

In earlier stages of the design process, we proposed the use of wearable activity trackers to measure physical activity and sleep. However, among the experts at the meeting, there was some consensus that long-term compliance rates for these devices are low, so we are unlikely to get comprehensive data collection for more than a few weeks at a time. This problem is likely to be exacerbated by our need to strictly control the feedback that participants receive. It may therefore be more efficient to do intermittent short-term activity tracker experiments

that are associated with specific motivational strategies, rather than to attempt to instrument the entire study population for years at a time.

It was noted that many kinds of data collection may be impossible across the entire study population, either because the data collection technology may not be stable across years or because of the high cost of data collection. When that is true, it may be advisable to design methods for intermittent data collection. It might, for example, be interesting to deploy advanced physiological monitors or other tools for 3-week periods to randomly selected study participants. One example of this would be to deploy environmental toxin monitors for 3-week periods randomly throughout the study population, such that 50 subjects are monitored in this regard at all times.

It was also noted that unstructured video and audio data might be banked during study enrollment. This data could be very easy to capture and bank, leaving open the possibility that future data analytic tools would make these datasets valuable.

Finally, it was also acknowledged that the overall study would benefit if independent investigators with their own funding were offered the opportunity to use new data gathering technologies within the study frame. A cardio-vascular study group, for example, might be offered the opportunity to recruit subjects for wearable EKGs for a limited period. The study management team will have to include a mechanism for reviewing proposals of this kind.

Focus Groups

Objectively, the proposed measurements span the spectrum of physical and personal invasiveness. However, opinions about what constitutes an invasion of personal privacy are quite diverse and idiosyncratic – some are willing to reveal detailed drug taking behavior, but recoil from discussing political party registration (even though this is a matter of public record), others feel that web search history should be more protected than location information, and in either case there are people with opposing views. It will be important to broadly sample opinions from our potential subjects about what they perceive as being invasive, and why, so that we can formulate strategies for addressing these concerns with our Public Outreach and Privacy Councils. We propose to implement focus groups in the coming months to begin this process.

Measurements

Proposed Data Collection Process for Subjects

1. Recruiter identifies candidates
2. Initial contact w/candidate (marketing materials)
3. Pre-screening by phone
4. Data collection process begins:

<u>Pre-enrollment (1-4 hrs, web based, in pieces, about a month prior to intake)</u>

- Current technology use
  - Apps (phone,computer)
  - Phone (model, OS)
  - Wearables
- Psychology questionnaires
- Download location tracker app to phone to gather a month's worth of data

Enrollment (1 full day, w/lunch break = ~6 hrs)

- Physical sample collection
- Medical history
- Review of questionnaire data (fill in any missing data)
- Remaining psychological testing
- Video capture of subjective measures
- Semantic labeling of prior month's location data
- Tech installation (including enrollment in any online services)
- Full consent process (staged form necessary for the pre-enrollment)

Forward going weekly/monthly/bi-annual pings to complete surveys

- Time budgeted to avoid overwhelming subjects
- Gamified – appropriately targeted to kids and to adults (e.g. social network construction instrument)
- Potentially to include small scale experiments for additional data collection (approved by board of directors, but not paid for by HUMAN project)- Potential examples include activity trackers or exposome silicone bracelets

Annual or biennial re-capture

- Return to enrollment center or use mobile home for follow up
- Physical exam

Life-triggered events

- Triggers include
  - Answers on the Traumatic Life Events Questionnaire
  - Large changes in the pattern of location data
  - Large changes in social network
- Triggers a specific data collection protocol to follow up

Data collection considerations:

- Incentives for each of these collection events
- How much feedback do we provide (do we tell people their weight and BP?)
- We envision stockpiling large amounts of data for future analysis because technology is improving rapidly (e.g. audio/video) and prices dropping (e.g. genetics). This raises some storage issues for banked data (both biological and digital):
  - Consent
  - Structured vs. Unstructured
  - Requirements for actual storage
- FOCUS GROUPS to determine what people consider invasive measurements (e.g. browser history, search history, etc.) and their comfort with long term storage of banked raw data

What We Should Measure (core study population ~3000 families)

NB: physical capture must be done at intake, all other measurements will be part of pre-intake or the ongoing survey budget.

<u>Initial demographics:</u>

- Examples for all domains

<u>Physical capture for health profiling – extra samples stored for later analysis:</u>

- Blood

    - Chemistry and blood count
    - Toxicology
    - A1C (diabetes)
    - CRP (heart disease)
    - Liver enzymes
    - Hormones
    - Some for storage
- Genotype (some sort of sample for storage, genotyping performed later as costs drop)
- Urine
- Saliva
- Microbiome (gut & skin)
- Hair
- Blood Pressure
- Heart rate/ Pulseox
- Temperature
- Height & weight

Psychological assessment (web-based whenever possible):
- Neo-PI (particularly conscientiousness)
- Locus of control
- Risk preference
- Impulsivity/discounting
- Executive function
- Working memory
- Attention
- PANAS (positive and negative affect schedule)
- Beck Depression Inventory
- Psychopathology inventory
- NIH toolbox for behavioral measures (http://www.nihtoolbox.org)
- Locus of control
- Emotional reactivity
- Mischel marshmallow test
- Stop signal task
- Clinical evaluation (SCID – Structured Clinical Interview for DSM-IV)

Social network assessment:
- Social capital
- Social media
- Safety nets (including physical location)
- Self-reports of social network (including physical location)
- Nuclear family interactions
    - Video of a conflict task or just observation (at time of enrollment)

Communication patterns (based on cell phones – in all participants >10 yrs old):
- Email (as much as we can get from phones)
- Texts (semantic labeling or full storage)
- Search/browser history from phone
- Apps (from phones – whatsApp, Skype, etc.)

Location data:

NB: Based on cell phones – in participants <10 ys old, we use Bluetooth beacons (gamified) to track when children are near their parents

- Bluetooth
- Wifi
- GPS
- Semantic labeling of location history

Health:

- Full medical history (including mental health, alcohol, tobacco, and drug use)
- Diagnostic history from NY State
- Hospital/physician medical records
  - Structured data (e.g. Text reports) for everyone
  - Unstructured data (e.g. Doctor's notes) as available – will try to partner w/Health and Hospitals Corporation, Sinai affiliates, Columbia affliates, and NYU affiliates
- Corporate medical/school medical records
- Medicare/Medicaid records
- Biennial re-capture of physical samples and updated medical history
- PatientsLikeMe
- Physical activity (via phone – maintenance of wearables may be too difficult)
- Self-report of weight and height (for children)
- Triggered detailed follow up on any catastrophic events

Education:

- School records: grades, standardized tests, absences, progression, IEP
- School description
  - DOE statistics for publics
  - Scrape ISAAGNY and InsideSchools database info into profile
  - Ask parents/children over 10 yrs old
    - Structured data (e.g. Does your school have a metal detector)
    - Unstructured data (interview)
- Delinquency inventory (for children over 10 yrs old)
- Dweck efficacy questionnaire
- After school and summer school programs (density and price)
- Teacher name – could we survey teachers to ask what % of kids do homework?
- Transcripts w/class names
- College prep programs (public or private)
- College
  - Profile of college (university, community college, etc.)
  - Transcripts
- Preschool programs
- Post-high school outcomes (college, armed forces, apprenticeship)

Work:

- Work history
  - From earliest work experience (part time work, after school jobs)
  - Total compensation (salary, variable income, fringes, and benefits
  - Promotions and transfers

- Parental education/work history
- Self-reports about work
  - Self-reports about work
  - How interesting is your job? Is it your calling? Work-life balance? Satisfaction? Autonomy? Work hour stability? Do you get paid for the hours you work? Harrassment? Comparison to family and friends jobs? What are your professional aspirations? Job security? Commute? Underemployment?
  - Ask families about their other family members work situation/habits
  - Who do you employ in your household?
- Work network (co-workers, supervisor)
- Licensure exams
- Union membership
- LinkedIn, Glassdoor

Financial assessment:
- Credit history (and consent for future reporting)
- Financial transactions (register for Mint)
- Wealth assessment
  - Home value (PLUTO, maybe Zillow)
  - Savings (regular and retirement)
  - Automobile
  - Other assests
- Insurance
- IRS data
- Social Security data
- Robin Hood Foundation Checklist for poverty
- Payment history on utility bills, and all bills going forward
- Liabilities
  - Including family members and friends
- Subjective sense of SES
- Banking information
- Other locations for financial transactions (payday loans, Western Union)

Socio-Political assessment:
- Voting data – frequency, party registration
- Large scale traditional survey
- Mass Media – what information are people getting about the world?
- Religion
- Moral behavior
- Volunteer activities/philanthropy
- Recreation activities

Known holes in the data stream:
- Cash transactions
- Barter economy
- Non-phone communications (Skype, gchat, etc.)

Browser and search history on non-phone devices (home/library device)

The Study Frame

The core group (N=~3000 families)
All of the above measurements

The hyper-study group (N = ~ 300 families)
(potentially a subset from the Quantified Community)
All of the core group measurements, plus:

- Magnetic Resonance Imaging
    - Highest possible resolution structural scan
    - DTI
    - Resting state
    - Functional scan w/passive image viewing (à la Todd Heatherton)
- EEG
- EKG
- Sleep study
- Detailed interviews (7UP style)

The crowdsource group (N=~300,000 people)
As many of the measurements as people are willing to participate in

- Survey games
- Sharing financial or health history and/or current health or financial data

Location tracking (w/map annotation)

Meeting Attendees

Nadav Aharony
Product Manager
Google

Dennis Ausiello
Jackson Distinguished Professor of Clinical Medicine
Director, MD/PhD Program
Harvard Medical School

Hannah Bayer
Chief Scientist
Institute for the Interdisciplinary Study of Decision Making
New York University

Jeanne Brooks-Gunn
Virginia and Leonard Marx Professor of Child Development and Education,
Teachers College and College of Physicians and Surgeons, Columbia University
Co-director, National Center for Children and Families
Co-director, Columbia University Institute for Child and Family Policy

Alin Coman
Assistant Professor, Department of Psychology
Princeton University

Paul W. Glimcher
Julius Silver Professor of Neural Science, Economics and Psychology
Director, Institute for the Interdisciplinary Study of Decision Making
New York University

Megan Huth
Research Scientist, People Innovation Lab
Google

Jennifer Kurkoski
Research Scientist, People Innovation Lab
Google

David Lazer
Professor in Political Science and Computer and Information Science, Northeastern University
Visiting Scholar
Harvard University

Kevin Ochsner
Professor and Director of Graduate Studies
Department of Psychology
Columbia University

Alex 'Sandy' Pentland
Toshiba Professor of Media Arts and Sciences
Director, Media Lab Entrepreneurship Program
MIT

Roberto Rigobon
Society of Sloan Fellows Professor of Management
Professor of Applied Economics
MIT

Benjamin Shiller
Assistant Professor of Economics
Brandeis University

APPENDIX C - 2

STUDY FRAME DESIGN ADVISORY COUNCIL WORKSHOP SUMMARY REPORT

December 12, 2014

Executive Summary

The Study Population Design Advisory Council met on December 12th, 2014 in Santa Monica to discuss options for designing and managing the proposed study population. The Council discussed all aspects of the study population, from the design of the study frame and recruitment of participants to strategies for long-term maintenance of the population and future replenishment of the study population. The highlights of the resulting design are described in the text below and a more detailed account that follows this summary.

Study Frame Design: A master study frame built from a number of separate components was deemed the best option for sampling the population of New York City in a statistically representative manner. Council consensus was to use physical addresses as the core of the study frame, relying on New York City's PLUTO database for frame construction. Because geographic databases may inject systematic biases into the study frame, a statistical model that also incorporates birth records, social security records, Medicare and Medicaid records, and school records to compensate for these biases will be constructed as part of the frame-building exercise. The statistical model will be a core asset maintained on an annual basis.

Although a single cohort (organized around statistically representative "seed" individuals) will be generated, that cohort will systematically oversample three groups: Children ages 1-3, Children ages 5-9, and post-retirement Elders. A member of the study staff with expertise in statistics and/or demography will work closely with the members of the Study Population Design Advisory Council and other experts to build the master study frame and the statistical model of the city during the next 12 months.

Study Pool Creation: A total pool of about 10,000 participants will be built by sampling from the study frame. In order to examine the influence of families, relatives, and residential groups, we propose to begin by representatively selecting approximately 2,500 individuals or "seeds." Once the model and a random process has been used to select a "seed," the selected individual and all of his/her co-residing family members would become study participants – the roughly 10,000 people that are fully instrumented and studied in extreme detail. These groups are referred to as "residential groups." Information about the extended, non-residential family network will also be collected, but data about the entire family network surrounding each study network group will be much less extensive. Our definition of a residential family rests on the current federal definition.

Strategies for Recruitment and Retention: All incentives have treatment effects and may influence the subject population non-uniformly. However, the problem induced by incentives is unavoidable, so our goal is to choose a set of incentives for recruitment and retention that minimize these effects. Incentives will likely include some combination of financial compensation, self-information, and information about the study results. During the next 18 months we will complete a focus group exercise and develop a series of pilot experiments to determine

the set of incentives that will be sufficient to insure effective recruiting without introducing unnecessary treatment effects.

Attrition: Unlike more traditional surveys, it may be possible to avoid the complete loss of participants from the subject pool if consent at enrollment includes permission to continue to collect "shed" data (data that do not require any action by the subject such as school records, Medicare records, and tax records), even after subjects stop answering questionnaires or providing location data. For this reason we identify two classes of attrition for the study: *Full Attrition* and *Partial Attrition*. Our goal is for a very small fraction of subjects to undergo full attrition and for this to occur only when subjects actively choose to remove themselves from all aspects of the study.

Out-migration: Although all recruitment will take place inside the geographic boundaries of New York City, we recognize that some subjects will migrate out of the city as the study progresses. We identify four classes of outmigration: 1) *Suburban Out-migration*: These are subjects who move immediately outside of the city. They will be retained in the study and followed like all other study participants. 2) *Temporary Out-migration:* These are subjects who leave the area temporarily or on a regular temporary basis. Elders traveling to warmer climates for the winter or students traveling to college are examples. These subjects will be fully retained. 3) *National Out-migration*: These are subjects who leave the city for other locations within the United States. These subjects are retained, as much as is feasible. All passive and non-physical data collection methods continue to be used on these subjects. 4) *International Out-migration*: These subjects are assumed to be lost to the study. Thus, only those people who leave the country will be viewed as completely lost from the study.

Subject Replacement: Subjects who choose to formally leave the study, out-migrate nationally, or out-migrate internationally will be replaced with new subjects, as to maintain the representative structure of our study population. These replacements will be selected based on the underlying study population model, which will be continuously updated as the underlying demographic population of the shifts.

Addition of New Subjects: New participants will also be added to the study population via births of children to seeds of study networks – these children will become the seeds of their own networks.

Future Waves: It is assumed that future waves of subjects, likely based exclusively on seeds ranging in age from 1-3 years old, will be considered in future years.

With these basic design elements in-hand, we can now work towards specifying the details of the study frame and methods for recruitment and retention of the study pool in the development of the CD1 and CD2 documents.

Detailed Summary

Key Discussion Points

Study Frame Design

The goal of the study is to examine the lives of a representative sample of roughly 10,000 New Yorkers, with a specific oversampling of three groups: young children, pre-teenagers, and the elderly. This will be achieved by drawing a random sample of the population from a study frame that represents New York City, oversampling our populations of interest. This raises the question of what frame the sample should be drawn from.

The advantages and disadvantages of several potential study frames were discussed, including addresses, birth records, social security records, Medicare and Medicaid records, and school records. Birth records, to take one example, provide the most accurate sample of infants, but provide a less representative sample for all other age groups. Social Security is similarly useful specifically for the elderly, and education records are especially effective for school-aged children.

To achieve our goal of sampling from the entire city in a representative manner, the Advisory Council proposed that we combine several potential frames to create a master study frame. The core of the study frame will be residential addresses in New York City, accessible through the Primary Land Use Tax Lot Output database (PLUTO). For every property tax lot in the city, PLUTO contains information ranging from census tract identity to the number of residential and non-residential units in multi-unit buildings. Note that each apartment building or condominium complex only appears once in the PLUTO database (though the total number of units in each building is listed). PLUTO thus serves as the primary study frame. We would then combine this with other records using a statistical model of the city to form a *Master Study Frame*.

Building the Master Study Frame will not be trivial and will be best accomplished by a demographic professional working under the supervision of the Advisory Council. This individual will take primary responsibility for building and maintaining our statistical model of the NYC population. This model, anchored to the Master Study Frame, will form the basis for recruiting subjects to the study.

In light of the need for expertise in statistics and demography, the study will have to recruit at least one expert in this area (which is not yet well-represented at the level of the Council) to join the Advisory Council and to help oversee and evaluate the work of the demographic professional(s) employed by the study during the CD2 phase of the project.

Building a Study Population

We could, in principle, choose 10,000 randomly sampled individuals from the Master Study Frame to build the study population. However, we hope to increase understanding about the influence of family and residential co-habitants on the lives and biology of our participants, particularly when those participants are children and elders. The HUMAN project is not a family-based study, but to gather data on the role of family and genetic relatives in participants' lives biology, we propose to include at least a subset of each participant's family in the core subject pool who we study extensively.

We therefore propose to choose roughly 2,500 sample individuals or "seeds" based on the Master Study Frame and then to build the study population outwards from them, much in the way that the Health and Retirement Study chooses an individual and includes both that person and his/her spouse in the study. Choosing whom to include in the core subject pool requires balancing the desire to include all individuals who strongly influence the seed individual or are related to him/her against the practicality that we cannot possibly include everyone with whom the seed spends a large amount of time, as some will have a more transitory role in the seed's life (for example, nannies or teachers). To balance these issues, the Advisory Council proposed that we build small networks, basically residential families, around our seeds and include all of those residential family members with the study cohort. Family members, defined by either biological or legal relationships, are the people who are most likely to contribute substantial long-term influences (genetic and/or environmental) on our seeds. It is for that reason that the Study Cohort is built around the residential families of our seeds.

Thus, we propose that once a seed individual is identified from the Master Study Frame, the selected individual and all of his/her co-residing family members would become study participants – the people that we fully instrument and study in extreme detail. We call this group a "residential network group." Note that we define co-residence as living together at least 5% of the time, and family members as people related by birth, marriage, or adoption as based on the US legal federal definition of family. The seed and all members of the seed's residential group will have to agree to participate in the study in order for any of the members of the group to be enrolled.

In some cases this definition would result in a rather large residential network: for example, a child of divorced parents with joint custody would be the seed of a network that included his/her biological parents, as well as any step-parents, step-siblings, or half-siblings who live with him/her. Conversely, a residential network may contain only a single person, if an individual lives alone. Note that the choice to define the residential network as a group of people related by birth, marriage, or adoption and residing together is consistent with the definition of a family used by the census, but more restrictive than current federal government definitions used to determine sick leave eligibility (which includes as a relative "any individual related by blood or affinity whose close association with the employee is the equivalent of a family relationship.") We note that roommates who do not meet the federal definition of family members will NOT be enrolled in the study when a seed is identified. The children of a seed living at college would, however, be enrolled as study subjects if s/he met the minimum residential time requirement. We expect that we would need to start with about 2,500 seeds in order to fill the study with 10,000 people.

Although our definition of the study group rests on a restrictive definition to determine eligibility, our intention is to collect extensive information about the extended, non-residential family networks of our subjects. This information can be tremendously important for a number of reasons. Familial data substantially increases the power of genetic analyses. For this reason, we propose to gather genetic information, not just about our subjects, but about their non-residential family members as well. Many adults have close relationships with family members (siblings, parents, extended family) and gathering information about these relations could provide substantial novel insight. It is not possible to study these very large networks of non-residential family members at the same level of detail as in the core study subject group (particularly for family members who reside outside of New York City), but using extended data collection, we may still be able to learn about the important influences of non-custodial parents or children who provide care to parents with whom they do not reside.

Strategies for Recruitment and Retention

Recruiting participants will be a critical challenge for the success of the study. Among the most important recruitment strategies will be to build (or borrow) institutional knowledge about recruitment. At the ISR, there is a training program in which experienced recruiters share their wisdom with new hires. However, even so, there are always some recruiters who are particularly talented, and these individuals should be deployed in difficult situations.

Like many other surveys, recruitment will be done by employing the standard *Dillman Method* – and of course this means we will have to offer incentives just to get in the front door. However, once people are enrolled, the Advisory Council believes that we will likely not have to provide as much in the way of incentives to retain subjects. Transportation may be part of the enrollment incentive package – sending a car service to transport people to the enrollment center is likely to dramatically improve the rate of people showing up. Communications strategies will need to be calibrated to reach the all of the audiences that we hope to reach. This will certainly involve communication in Spanish and, to a lesser degree, Chinese languages.

It may be difficult to recruit children into the study. Those who work in education have raised the issue that they typically have more success with community-based recruitment (e.g. going to street fairs). However, this is not compatible with our study frame, so we will have to determine whether we will need to find adjunctive approaches to reach out to populations that may be less comfortable participating in a research study. To this end, we propose to run a pilot experiment in recruitment towards the end of the final (CD2/CD3) design stage in roughly 18 months, to ensure that our recruitment plans will provide a sufficient yield to fill the study population at a reasonable rate.

Incentives for recruitment and retention are a critical part of the study design, but still require substantial thought before they can be further specified. Some incentives, like annual Christmas cards and a web portal that contains findings that have come out of the study written for lay people, provide minimal interference and engender a lot of good will for little money. These will certainly be features of the study. However, more intensive incentives, in the form of money, services (such as health care or cell phone contracts), or information, all have treatment effects. Even worse, these treatment effects interact with the different demographics of the subjects. It will therefore be paramount to design a set of incentives that provide the minimum monetary compensation and information required to recruit and retain participants. Once selected, the effects of these incentives on our subjects will have to be explicitly modeled by our econometric staff. However, we need to recognize that even asking people questions as part of a survey is creating a treatment – there is no "pure" sample. Detailed models of these treatment effects will have to be developed in the CD2 design stage.

Below are specific examples of the kinds of differential treatment effects and challenges we face:

- Providing smartphones to those that don't have them
    - This is a treatment for only some participants (most likely to be prevalent in the low SES and elderly).
    - What if people (particularly kids or low SES participants) sell or trade the phone we provide? Providing cellular contracts with phones may ameliorate this issue.
- Elders will have to be incentivized to switch from their current technology
    - We could set up a way to automatically trigger a call for help based on activity patterns or send blood pressure info to a doctor as an incentive.
    - We might want to use Bluetooth beacons (on a walker or bracelet) as a backstop measure for technology resistant subjects.

Attrition

There are a number of approaches to prevent attrition, including the use of retention specialists, institutional knowledge for keeping people enrolled, and re-contact (though not so frequently that it becomes harassment). There is also a cost benefit analysis to be done – a large fraction of the budget is spent to retain a relatively small fraction of the participants (~10%), so some decisions must be made about whether there are resource limits in preventing attrition.

However, as the study relies extensively on automatic data collection, it may be possible to avoid having people entirely disappear from our data logs, even if they choose not to actively participate in the study after some point. To make this possible, at initial consent it will be essential to obtain consent to continue passive data collection about our subjects even if they choose to stop engaging in active responding. Ideally, we will be able to track subjects even after they stop actively responding via Social Security records, Department of Education records, SPARKS data, IRS data, and Veterans Administration data, to take several examples. Of course, subjects must have the option to entirely opt out of further data collection if they so desire, but it is assumed that this form of absolute attrition will be rare. In this way, we should be able to continue to collect data about our participants,

even if they stop actively providing data (e.g. Answering surveys or providing location data via smart phones or other automated devices).

A number of issues related to attrition also arose which require further study and attention. We will need, for example, to budget time and money for re-consenting participants as they turn eighteen years old. This will be critical as these subjects may be particularly prone to withdraw from the study. People will also change life circumstances, going to places where it will be difficult to get active involvement in data collection (e.g. colleges, nursing homes, or prisons). Some of these demographic changes may be easier to plan for (e.g. it may be possible to catch college students if they come home for school breaks), but this is another area where it will be necessary to do a cost-benefit analysis to determine how much should be done to track people under difficult conditions.

Out-migration

The more common life change for which we will absolutely need to be prepared is a move out of New York City. While demographic data suggests that a representative cross-section of the city will yield subjects who overall move very little, higher income groups, in particular, do move. For the study's purposes, moves can be divided into four categories, each of which we address with a different strategy:

- *Suburban Out-migrants*: People still in the New York metropolitan area – keep in study (they probably still come to the city sometimes, so should not be too hard to get for physical follow up). These subjects will not be replaced.
- *Temporary Out-migrants*: People temporarily out of the NYC area (snowbirds, college students) – keep in study, try to follow up when they are in the NYC area (summers, etc.) These subjects will not be replaced.
- *National Out-migrants*: Still in country – collect shed data (e.g. Gov't records, keep pushing surveys, and collect cell phone data). While data continues to be collected, these subjects will be replaced.
- *International Out-migrants*: Out of country – gone from survey, these subjects will be replaced.

Adding New Participants

As people die or leave the study, the study population will need to be refreshed. New participants will be selected to replace those who are lost based on a dynamic model of the NYC population via the Master Study Frame. In this way, the study population will continue to reflect the ever-changing population of New York City as much as possible.

Although we expect that residential groupings of our study participants will change over the course of the study, anyone who becomes part of the study through a seed individual will be retained in the study population. For example, if spouses in a residential network divorce and move to different residences, both will remain participants in the study. In addition, if they re-marry, their new spouses, children, and any co-residing stepchildren will also become part of the study population. The study population will also expand over time, as any baby born to a study seed will also become the seed for a new network.

There are not yet plans for a second wave of the study. However, targeted fundraising could be used to support additional cohorts. These efforts might address the need to study particular populations of interest or the implementation of new technologies. We anticipate considering a second wave 5-10 years after study initiation.

Additional Study Design Issues

There is still considerable disagreement about the use of triggered data collection. For example, if a participant reports the occurrence of a major life event on a regularly scheduled survey, an automated protocol for collecting additional data could be initiated. In this way, the study could increase the density of information gathered during particularly interesting intervals in the lives of participants. This additional data could be incredibly helpful in understanding what happens when a baby is born or a person becomes unemployed. However, important issues were raised about whether such non-uniform data sampling was problematic from a statistical or a privacy perspective. There was some concern about what might be appropriate triggers and whether such additional data collection would be intrusive at times when participants would be particularly sensitive to intrusion.

Another important aspect of study design that remains unresolved is the use of a crowd-source population. It is appealing to envision deploying the measurement instruments designed for the study in a very large group of interested citizen scientists. Data from an additional 300,000 New Yorkers would substantially increase the power of our measurements. They would also provide a test bed for proxy measurements designed based on effects observed in the core subject pool. However, it is important to remember that this would be a sample of opportunity, and selection bias could substantially limit the generality of any observations. Thus, we will need to evaluate the cost of a crowd-source subject population both in dollars (where it could be a significant drain) and in public relations (where it could be a bonus, but only if managed correctly).

Although procedures during the intake procedure would help to identify mental health problems in participants at initial screening stage, it is also expected that additional mental health issues will arise over the course of the study – the frequency of such diagnoses could be as high as 1 in 5 young people. These subjects may require specific assets from the study after the onset of mental illness. We also expect that a subset of the elders in the study will develop dementia. Procedures will need to be developed for handling the consent (and possibly re-consent) process in these situations, as well as for facing attendant difficulties with data collection in dementia patients.

Policies will also need to be developed to identify and resolve conflicts between self-reports and administrative data. In many cases, it may be possible to use related data that has already been collected to address this issue. For example, financial transaction data and location data could be compared to self-reports about how many hours a participant spent working a part-time job. Such comparisons might also be useful in identifying fraud, but other procedures might be necessary as well. This raises the issue, as yet unresolved, about what conditions we would remove someone from the study for fraud or misrepresentation.

There will be a tremendous amount of observational data collected during the study, but one way of increasing inferential power would be to look at the effects of policy changes (particularly those limited in either time or location). In order to capitalize on these "natural experiments" the sample would benefit from increased power where there are policy changes (e.g. Children for education policy, geographically for other kinds of policies), and of course, policy changes must be recorded in the database.

This data set could be overwhelmingly complicated for some scientists, particularly those with limited experience in analyzing large survey data sets. One way to improve the accessibility of the data would be to create "training wheels" for data analysis. This could include subsets of the data, summary statistics that would be of particular interest, and other data products that would lower the barrier to initiating an analysis of the data set. For example, RAND has done some work like this for using Health and Retirement Survey data and the National

Longitudinal Survey staff created online tutorials for improving the ease of using the data from the NLS97 data set. Developing a "path to expertise" for scholars will be a critical feature of the study's professional communications strategy.

The current set of proposed measurements provides a relatively complete picture of the objective socio-economic status of our participants. However, recent work suggests that subjective perceptions of socio-economic status can be quite divergent from the objective measures, and that the subjective perception has a much bigger effect on outcomes. This suggests that subjective SES will be an important quantity to measure. One method for measuring subjective SES is to show participants a ladder of SES levels and ask them to put themselves on what they perceive as the correct rung.

One potentially important source for student records is the National Student Clearinghouse (http://www.studentclearinghouse.org/). This resource contains a nearly complete database of student enrollment and degree records. It provides reporting services and regularly collaborates with institutional researchers.

Meeting Attendees

Hannah Bayer
Chief Scientist, Institute for the Interdisciplinary Study of Decision Making
New York University

BJ Casey
Director, Sackler Institute for Developmental Psychobiology
Professor of Developmental Psychobiology
Weill Medical College of Cornell University

Miyoung Chun
Executive Vice President of Science Programs
The Kavli Foundation

Brian Elbel
Associate Professor of Population Health and Health Policy
NYU School of Medicine

Paul W. Glimcher
Julius Silver Professor of Neural Science, Economics and Psychology
Director, Institute for the Interdisciplinary Study of Decision Making
New York University

Arie Kapteyn
Executive Director, Dornsife Center for Economic and Social Research
University of Southern California

Kenneth M. Langa
Professor of Medicine
University of Michigan

Matthew D. Lieberman
Professor of Psychology, Psychiatry and Biobehavioral Sciences
Director, Social Cognitive Neuroscience Laboratory
University of California, Los Angeles

Christopher Martin
Science Program Officer
The Kavli Foundation

Kathleen McGarry
Department Chair, Professor Department of Economics
University of California, Los Angeles

Derek Neal
Professor
Department of Economics & The Committee on Education
University of Chicago & NBER

Sharif Taha
Science Program Officer
The Kavli Foundation

Paul Thompson
Professor of Neurology, Psychiatry, Radiology, Engineering & Ophthalmology
Director, NIH ENIGMA "Big Data" Center of Excellence
Associate Dean for Research, Keck USC School of Medicine
Director
USC Imaging Genetics Center

APPENDIX C - 3

PRIVACY AND SECURITY ADVISORY COUNCIL WORKSHOP SUMMARY REPORT

January 20, 2015

Executive Summary

The Privacy and Security Advisory Council met on January 20th, 2015 in New York City to discuss options for designing the privacy and security policies for protecting the data and subjects associated with the Kavli HUMAN Project. We discussed all aspects of the data management, from privacy policies and consent procedures to strategies for protecting the data from malicious attack and ensuring disaster recoverability. The highlights of the resulting design are described in the text below and a more detailed account that follows that overview.

Third Party Requests: Any valuable database will be the target of third party requests, and the HUMAN project data could be attractive to a number of entities, including, but not limited to, government officials, police officers, and divorce lawyers. However, many participants would be reluctant to share their data if it was accessible by such parties. In order to protect the data from third party requests it was agreed that we must apply for and obtain a **Certificate of Confidentiality for Health Research** (COC) from the National Institutes of Health (NIH). COCs are granted in order to protect identifiable research information from forced or compelled disclosure. With such a certificate in hand the study cannot be subpoenaed, granting strong protection of our subjects' privacy. Many federally funded research projects on sensitive topics such as drug abuse have received COCs, and the U.S. Health and Retirement Study has also been granted one, suggesting that this will be a viable solution to the problem of third party requests. Gathering more information on the willingness of NIH leadership to grant a COC will be a priority item for the next 6 months.

Data Sharing and Governance Policies: A primary goal of the HUMAN project is to make the data collected under its auspices an open resource for scholars. However, there are substantial privacy and security risks associated with data sharing, so policies will need to guard against such breaches. Policies should be as permissive as possible, but the beneficial outcomes of analyses must be weighed against ethical concerns, de-identification risks and the possibility of adverse legal bias or oppression for each potential project. We must also consider potential public relations ramifications, fallout from which could endanger the feasibility of the project by compromising relations with participants and funders.

Data sharing policies will therefore have to be developed to guide decisions about who is allowed to use what data. These decisions will need to be made on a case-by-case basis, following the evaluation of individual use case proposals. (We anticipate early drafts of these policies for the CD2 document.)

The *Research Proposal Evaluation Committee* will thus play a crucial role in the interpretation and implementation of data sharing policies, analogous to the role an institutional review board plays in ensuring the protection of human subjects in academic research. Accordingly, this committee must have representatives who can speak for the diverse interests of those who are involved in the collection and use of the data: individuals from the research

community (both non-profit and for-profit institutions), the study staff, and most importantly, the participant population. A full design for the Research Proposal Evaluation Committee will need to be created for the CD2 document.

It will be important to keep representatives of the participant population involved in the design data governance and privacy polices more generally as the study progresses. To ensure their representation, we suggest that in addition to their presence on the Research Proposal Evaluation Committee, there should also be a representative from the participant population on the Privacy and Security Advisory Council. The representatives to these committees should be drawn from a participant committee that is composed of a diverse sample of study pool participants. These individuals will help form a consensus that represents and protects the interests of the participants in policy decisions.

Recombination of HUMAN Data with Federal Administrative Data: One way of increasing the power of our data would be to combine it with records from the federal government, as well as with data from social programs administered by the NY state government. However, each of these entities has their own privacy regulations, so it will be necessary to comply with all relevant policies, particularly with regard to data recombination. Recombination of our data with detailed census data, though powerful, will be a special consideration, as census data is only accessible at a Census Research Data Center. Thus, in order to do the recombination, de-identified data would have to be extracted from our database and then moved to the RDC servers for analysis. This would require consent from participants to give our identified data (including social security numbers) to the government for linking. Opportunities for recombination thus require careful consideration and will likely have to be addressed during the CD2 and/or CD3 process.

The Three-step Consent Process: The Council envisions the main consent process as a three-stage process. First, a video will be presented which describes the basic issues that require consent, as well as insights from secondary use cases for the data. The video will be divided into segments, and after each one a member of the consent staff will stop the presentation to administer comprehension checks and facilitate a discussion about the issues raised in the video. After the video, participants will be presented with the opportunity to provide oral consent. If they do so, the third step will be the presentation of the long paper form for participants to sign. This paper form will contain all of the legal language necessary for written consent, but all the information in the form will have been covered during the video and discussion process. Thus, we expect that the acquisition of written consent will be relatively straightforward and a somewhat incremental step of the consent process. The intake process will be scheduled no sooner than twenty-four hours after the consent process, so that participants have time to re-consider (and retract consent) before data collection begins. Although the consent process could take place in the clinical setting where data collection occurs, for a number of reasons it might be more feasible to do it in the homes of participants. This will, of course, require that each family be individually consented in their home by trained staff.

Ensuring the Protection of Children in the Study: A major challenge for ensuring informed consent in study participants is the collection of detailed data (particularly genetic data) from children. As required by law, parents will provide consent for their children to participate in the study and children will *assent* to their participation. When participants reach age eighteen, they will have to go through the consent process as adults and agree to continue to participate in the study. We expect to be able to do genetic sequencing as well as detailed educational and cognitive profiling of children, creating a trove of very sensitive data about these children. While parents may feel comfortable consenting for their children to contribute such data to the project, young children may not have the capacity to understand the implications of providing genetic data or long-term detailed data collection in general. It remains undecided by the Council whether it may be more ethically justified to give participants the option to withdraw their highly sensitive data (which would have been gathered over the past 1-18 years of the

study) from the database when they reach age eighteen. As this could substantially weaken the power of the study, further discussion of the ethical obligations and consideration of the policies of other longitudinal studies, such as the National Longitudinal Surveys (NLS) and the National Children's Study (NCS), will be necessary before making a final decision on this issue. A survey of the NLS and NCS policies on this matter will be undertaken within the next 6 months.

The Structure of Data Storage: We propose to implement a data platform similar to the data warehouse facility that NYU's Center for Urban Science and Progress (CUSP) is currently building. In a facility of this design, data goes through a staging process as it is ingested into the warehouse for storage. Then, when researchers want to run queries or perform analyses, *temporary* specialized "data marts" are created which contain only the relevant data sets. Data can only move in one direction through the data warehouse, so that unauthorized access to the warehouse cannot be achieved from the staging server or the data marts; this is controlled by both electronic and physical means.

Three features of the data marts enhance the security of the data. Each data mart is created for a specific project, and researchers have access to only the data mart(s) required for their own project for the amount of time required to perform the necessary analyses. After the researcher is finished, the data mart is removed by deletion (though the data itself always remains safely stored in the warehouse in its original form), so that it is not vulnerable to unauthorized access. The data marts themselves are thus heavily partitioned project silos.

Other security measures include integrity checking of the data that comes off participants' phones, computers and outside servers, as well as detailed network segmentation.

Access Controls: In addition to the protective structure of the data warehouse architecture, additional physical and electronic access controls will also be necessary to ensure the security of the HUMAN project database. Physical access controls include securing the server rooms and limiting who has access to those spaces. Electronic access should generally be granted on an as-needed basis, and broader access limited to a relatively small number of people who have passed strict federal-level background checks and regulated by role-based access privileges. Even with careful administration of access privileges, it will be important to implement additional steps to protect the data from malicious insiders (or someone stealing the credentials of an insider's account). Measures such as login monitoring and behavioral monitoring of administrators can be very powerful and are standard in most high security data environments. However, these automated alerts for unusual patterns of behavior work best for long-term employees who can be easily and closely monitored, and for whom there are clear expectations of behavior. If we expect to allow scholars to access the database directly, it may be more difficult to use these approaches with that group of individuals.

Credentials for researchers will be considered in the context of the evaluation of data sharing proposals. Even after substantial vetting, external researchers, particularly those working off site are likely to be the greatest vulnerability to the system. Training in security and data management prior to any data access will be critical to ensure that researchers comply with all policies and regulations. An additional approach would be to encourage collaborations with staff researchers. Such collaborations would provide additional protection by limiting the number of people who can access the data, but could also add value for researchers (particularly those who are inexperienced with handling large, sensitive data) by offering an opportunity to learn from an expert. Alternatively, or additionally, in-house technical staff with expertise in data management could be employed with the sole job of assisting researchers in accessing data.

Special Protections for Sensitive Data: As is the standard protocol at CUSP and many other places, all data ingested into the database will be classified along the continuum from highly sensitive (and thus most stringently access restricted) to not at all sensitive (publicly available). This provides the opportunity to increase security measures for the most sensitive data without creating excessive barriers to the use of less sensitive data. This may mean adding additional layers for separating sensitive data, triple de-identifying it with different codes for different types of data, and using a 2-key system (i.e. 2 administrators required) to access the most private sets of data. Separating out sensitive data with different levels of security will be necessary at the level of both data storage and data marts.

Detailed Summary

Key Discussion Points

Privacy

The discussion was largely focused on privacy and security of the electronic database that will be associated with the study. However, it will be important to also consider the privacy and security of all stored biological samples. Currently we have proposed to store samples of biological material such as cell lines, genetic material, blood, urine, and hair. Although the protection of physical samples was not discussed in detail at this advisory council meeting, a separate plan will be necessary for biological samples and it must meet equally stringent requirements for privacy, security, and disaster protection as the plan that will be designed for protecting electronic data. Physical samples will likely be stored in a commercial facility. The security standards for physical sample storage should be developed for the CD2 document.

An issue that was raised in several different contexts was the separability of risks from privacy/security breach and public relations/trust problems. These are related – certainly the former will likely result in the latter. However, there are circumstances in which policies that meet legal requirements might still result in outcomes that weaken the trust of participants and/or the general public in the project. Therefore, it is critical to consider the consequences of all potential policies in light of both views.

Data access rights are still in flux, but the legislative environment is evolving quickly, and given the proposed strategy to provide only partial feedback as a way of controlling for treatment effects, it will be important to be transparent about the limitations we will place on participants' ability to access their own data. A new Consumer Privacy Bill of Rights and a new Student Digital Privacy Act (based on California legislation – Student Online Personal Information Protection Act) are expected to be unveiled relatively soon.

A number of data types are protected by specific government regulations: data from health care providers by HIPAA (Health Insurance Portability and Accountability Act), data from K-12 educational institutions by FERPA (Family Educational Rights and Privacy Act), and data from financial institutions by GLBA (Gramm-Leach-Bliley Act). Since the project is not run by any such entities, there is no requirement to be in compliance with these regulations (which are quite onerous). It was suggested that it would make more sense to be "in accordance" with such regulations, which would mean meeting those standards, but not bearing quite as heavy a burden. For example, this is how insurance companies function, allowing them to give representation and assurances to people that there is no additional risk beyond what they already experience.

It will be important to look out for the interests of the participants in the study as decisions are made about the use of study data as well as changes to the study over time. One important way to address this would be to add an advisory council made up of participants – a possibility also raised by the Subject Pool Advisory Council. Members of this committee would be chosen to reflect the diversity of the subject pool, and the chairs of that council would serve to represent the participant perspective to other decision making groups for the study.

A more detailed discussion of consent follows below, but one consent-related privacy issue is mentioned here. Among the plans for expanding the study, it has been proposed that investigators might initiate independently funded projects that capitalize on core participant groups. Additional consent processes would likely be necessary for such initiatives, but it should be noted that the study staff would have to manage these additional consents in order to protect participant identity.

Third Party Requests

Any valuable database will be the target of third party requests, and the HUMAN project data could be attractive to a number of entities, including, but not limited to, government officials, police officers, and divorce lawyers. These requests could be frequent and extensive, though it is difficult to gauge at this point. However, it will likely be necessary to allot budget for legal fees and staff for handling requests.

Several strategies for reducing vulnerability to (and the costs of) third party requests were proposed, but dismissed as providing insufficient protection. If the database were to be under the auspices of the federal government, it would be protected from civil litigation, but would still be accessible by FOIA requests. Moving the data offshore is also not a solution, because the standard of the law is that if you can access the data for scientific purposes, then you are obligated to access it upon legal request.

It is therefore best to limit stored data to what is absolutely necessary. For example, it would be ideal to remove identity data entirely from the database; this would provide protection from both subpoena and hackers. Aggregation would have a similar effect – once the data is mixed, data about any individual could not be extracted. However, neither of these are straight forward solutions for this particular project – it is difficult to envision a mechanism that could be used to continually accrue data without an identifier linked to real-life identity at some level. One suggestion would be to have a separate team to do all participant interaction and to keep this team away from the data so that there is a plausible separation. In this way, either team alone would not have the capacity to match data with identity, and the groups might not be legally compelled to cooperate if they were separate entities (although this latter condition was not certain). If a participant fingerprint was required for linking certain kinds of data (e.g. Re-identification), this might also strengthen the ability to withstand legal challenge, since the participant would be the only one able to provide the requested data.

However, a de-identification solution would still leave the data vulnerable to requests where the requestor already knew something about the records being requested (for example: "please provide all data pertaining to the individual who was at location X at time Y on date Z"). Separating the control of individual data stores for identity, location, genetics, and other sensitive data would provide the opportunity to put in place additional protections for such data sets. Such separation would certainly be useful for strengthening security measures, but is unlikely to be useful in blocking legal requests for data.

Given all of these constraints, the optimal solution for this project is to obtain a **Certificate of Confidentiality for Health Research**. From the COC guidelines:

Certificates of Confidentiality are issued by the National Institutes of Health (NIH) and other HHS agencies to protect identifiable research information from forced or compelled disclosure. They allow the investigator and others who have access to research records to refuse to disclose identifying information on research participants in civil, criminal, administrative, legislative, or other proceedings, whether federal, state, or local. Certificates of Confidentiality may be granted for studies collecting information that, if disclosed, could have adverse consequences for subjects, such as damage to their financial standing, employability, insurability, or reputation…

Certificates of Confidentiality protect subjects from compelled disclosure of identifying information but do not prevent the voluntary disclosure of identifying characteristics of research subjects. Researchers, therefore, are not prevented from voluntarily disclosing certain information about research subjects, such as evidence of child abuse or a subject's threatened violence to self or others. However, if a researcher intends to make such voluntary disclosures, the consent form should clearly indicate this.

The HRS has a COC, providing additional confirmation that the project would be of appropriate scope to request one. Discussion with HRS advisors seems sensible, and additional detailed information may also be found at: http://grants.nih.gov/grants/policy/coc/index.htm

Data Sharing

A primary goal of the HUMAN project is to make the data collected under its auspices an open resource for scholars, but balancing data sharing with privacy concerns is not trivial. Sharing with researchers from industry, such as pharmaceutical company employees, has additional potential for concern, particularly since it is critical to maintain independence from for-profit entities. Therefore, it will be important to ensure that strong governance policies are in place to guide decisions about data sharing.

Policies should be as permissive as possible, but the beneficial outcomes of analyses must be weighed against ethical concerns, de-identification risks, the possibility of adverse legal bias or oppression, and potential public relations ramifications, fallout from which could compromise the feasibility of the project by compromising relations with participants and funders. Each research proposal that requests use of the data will be evaluated with respect to these criteria - articulated in the form of data governance policies – by a committee including representatives from all of the stakeholder communities, the *Research Proposal Evaluation Committee*. This will include civil rights advocates, privacy experts, study participants, academic researchers, and corporate researchers. An oversight monitoring and auditing mechanism will also be necessary.

At this time, the Council advises that we not allow use of the data for commercial or marketing purposes, but there may be industry users who have substantial interest in using the data for health or education research. There was wide agreement that such a use fell well within the study's overall mandate. Nonetheless, as the research results would benefit for-profit corporations, it may make sense to charge these entities for use of the data. A tiered pricing model would take into account a number of user factors, for example the size of the organization and the number of data points or data stores being accessed. There are some existing licensing models of this type that we might look to for guidance. However, profit or commercialization of the database are not study goals, and would certainly have negative ramifications, so it will be critical to make clear that any money obtained from for-profit users will be reinvested in providing financial aid to researchers from non-profit institutions. This is also important because if the project does make a profit, there may be issues of property rights – i.e. what people should get out of their own data if we are profiting from it. During the CD2 phase a proposed fee structure and regulations for users of different types will need to be drafted.

Recombination of Our Data with Administrative Data

One way of increasing the power of our data would be to combine it with government records from the Census Bureau, the IRS, and the Social Security Administration, as well as data from social programs administered by the state government, such as SNAP, Medicare, and TANF. However, each of these entities has their own privacy regulations, so it will be necessary to comply with all relevant policies, particularly with regard to data recombination. As an aside, it is relatively easy for individuals to get their own SSA and IRS records released to them, but the protocols typically require that the data be sent to the individual, not to a third party. Some discussion with the HRS leadership may be necessary to determine whether it is possible to arrange bulk release of records.

However, recombination of our data with detailed census data will require a unique protocol, as such data is only accessible at a Census Research Data Center. There is already an RDC in Manhattan, at Baruch, though it might be possible to create one at NYU/CUSP. However, in either location, identified data would have to be extracted from our database to combine with census data on the RDC servers. Consequently, we would need to get consent from participants to give our identified data (including social security numbers) to the government for linking. During the CD2 and/or CD3 process we will carefully consider opportunities for recombination and outline a protocol for implementing such analyses while ensuring the security and privacy of our participants' data.

Consent

The Main Consent Process

Getting informed consent for this complex study will require a process carefully designed to ensure that all participants understand the risks and benefits of the research we hope to undertake. Given that there are analyses that we have not yet conceived (with equally unknown outcomes) it will require a careful balance between being as specific as possible and leaving open future possibilities. We expect that there will need to be additional consent processes as the study progresses, but we plan for an initial comprehensive consent process that will address the first five years of the study.

We envision this main consent process as a three-stage sequence. First, a video will be presented which describes the basic issues that require consent, as well as insights into secondary use cases for the data. The video will be divided into segments, and after each one, a member of the consent staff will stop the presentation to administer comprehension checks and facilitate a discussion about the issues raised in the video. After the video, participants will be presented with the opportunity to provide oral consent. If they do so, the third step will be the presentation of the long paper form for participants to sign. This paper form will contain all of the legal language necessary for written consent, but all the information in the form will have been covered during the video and discussion process. Thus, we expect that the acquisition of written consent will be relatively straightforward and a somewhat incremental step of the consent process. The intake process will be scheduled no sooner than twenty-four hours after the consent process, so that participants will have time to re-consider their participation before data collection begins.

There are a number of ways that the full consent process could be administered. One possibility would be to perform it at a clinical facility, as the first step of the intake process. However, this has a number of disadvantages – most importantly, it does not allow much of a waiting period before data collection begins, and logistically it would make it more difficult to efficiently schedule data collection during the intake process, since in order to enroll all members of a family, they would have to spend the first part of the day participating in the consent

process. Instead, we propose that the consent process be administered in the participant's own home. This may also help to put people at ease, as they will be in a familiar space. Staffing costs for this process will need to be estimated for the CD1 document.

Additional Consent Processes

The consent process will be administered for all members of the residential network at the same time, prior to intake. However, there may also be a need to get consent from people who are not members of the study. Non-residential family members who we hope to follow as auxiliary members of the study (doing less intensive collection of genetic and behavioral data) will need a separate consent process, and it may also be appropriate to obtain consent from care providers who spend substantial amounts of time in the home (i.e. nannies and home health care aides). As we plan to collect only metadata about communications partners, it has been tentatively concluded that it is not necessary to obtain consent from them, but as there may be a privacy issue (or the perception of one), it may make sense to provide some notice of disclosure during the communications to our participants.

We expect that the initial consent process will provide adequate protection for study participants in achieving the main study aims. However, over time, it may be advantageous to expand the scope of the study with the use of additional technology or secondary use cases. These situations will likely require additional consent from our participants. As described previously in this document, for the preservation of privacy, it will be necessary for the study staff to handle this process. For efficiency, it would be preferable if additional consent occurred at the biannual appointment for physical specimen collection, phasing in large scale changes over time, but an additional appointment with study staff could be arranged for urgent projects or for those based on a small subsample of the subject pool.

Ensuring the Protection of Children in the Study

A major challenge for ensuring informed consent in study participants is the collection of detailed data (particularly genetic data) from children. As required by law, parents will provide consent for their children to participate in the study and children will assent to their participation. At NYU, children twelve years or older are considered capable of providing written assent. When participants reach age eighteen, they will have to go through the consent process as adults and agree to continue to participate in the study. Given that the research here will not provide a direct benefit to participants, we note that it was the consensus of the Council that no child can be compelled to provide data of any type unless he or she freely assents. This is true even if the child's parent or guardian assents. So, for example, a 9 year-old child cannot be compelled by his/her parents to permit a blood draw for the purposes of the study. *This is a critically important limitation on data collection from children.* We therefore assume that invasive measurements such as a blood draw may not be possible in children or teenagers. However, we do expect to be able to perform genetic sequencing (which does not require blood), as well as detailed educational and cognitive profiling. These are measurements that will result in the collection of extremely sensitive data about children. While parents may feel comfortable consenting for their children to contribute such data to the project, young children may not have the capacity to understand the implications of providing genetic data or long-term detailed data collection in general. For this reason it remains undecided whether it may be more ethically justified to give participants reaching adulthood not just the opportunity to leave the study, but also the opportunity to withdraw their highly sensitive data from the database – even if that data has been gathered over 18 preceding years. Further discussion of the ethical obligations and consideration of the policies of other longitudinal studies (such as the NLS and NCS) will be necessary before making a final decision on this issue. Assessing these policies will need to be completed before the end of the CD1 phase.

Another issue for discussion is whether the Children's Online Privacy Protection Act (COPPA) is relevant to the study. COPPA pertains to websites collecting information from children under the age of 13. If a web-based authentication process is used for data collection, COPPA may apply. This regulation was issued by the FTC, and is largely about obtaining parental consent for any data collection, which is already incorporated into our procedures. However, it does entitle parents to request deletion of their child's data at any time, which could have serious consequences for the study. Further legal advice on this matter will be necessary to determine the implications for study policies prior to the completion of the CD2 phase.

Other Vulnerable Populations

Over the course of the study, we expect that some elders will begin to lose their mental faculties. For those that are legally declared incompetent, there will be an individual designated as having power of attorney, and this agent will be the person responsible for any decisions or consent requirements. However, there may be situations where study staff, through interactions with a participant, may perceive that mental function is compromised, though no legal measures have been taken. In these cases, a member of the study staff will need to identify an advocate or interested agent with whom they can consult. However, some policy will need to be set in order to determine the threshold for invoking this process. These policies will need to be defined and legally vetted prior to completion of the CD3 stage.

Over the course of the study, some participants will also go into prison, where active data collection will be difficult. However, it may still be possible to collect data passively (especially court and arrest data, which are a matter of public record), and certainly we would hope to be able to use any data on these people to understand the role of prior experience in post-incarceration outcomes. To that end, it will be necessary to put in place policies to address these issues, and to justify the importance of this research as it benefits the prison population as a whole, and more generally supports studies of recidivism.

Data Management and Security

Data Platform Structure

We propose to implement a data platform similar to the data warehouse facility that CUSP is currently building (Figure 1). Data goes through a staging process as it is ingested into the warehouse for storage. Then, when researchers want to run queries or perform analyses, specialized temporary "data marts" are created which contain only the relevant data sets. As indicated by the arrows in Figure 1, data can only move in one direction through the data warehouse, so that unauthorized access to the warehouse cannot be achieved from the staging server or the data marts. This will be achieved through both electronic means as well as physical means (simply cutting the pins on the connection cables and creating a "data fountain").

Three features of the data marts enhance the security of the data. Each data mart is created for a specific project, and researchers have access only to the data mart(s) required for their own project for the amount of time required to perform the necessary analyses. After the researcher is finished, the data mart is deleted (though the data itself always remains safely stored in the warehouse) so that it is not vulnerable to unauthorized access. The data marts themselves are thus heavily partitioned silos.

Some additional security measures will also be necessary because data will be coming directly from our participants' smartphones or other devices outside our direct control. In order to avoid contamination, it will be necessary to scan for malware and viruses prior to ingestion. The applications on the phones themselves will need to be written with a way of checking the integrity of the incoming data.

Network segmentation is also a critical feature of the system design. Security breaches are unavoidable, despite even the most stringent approaches, and splitting up the network can help improve security. For example, it may prevent an intruder from gaining broad access with stolen network credentials. Creating sub-networks or layers can also make the network more robust by limiting the effects of local failures on other parts of the network.

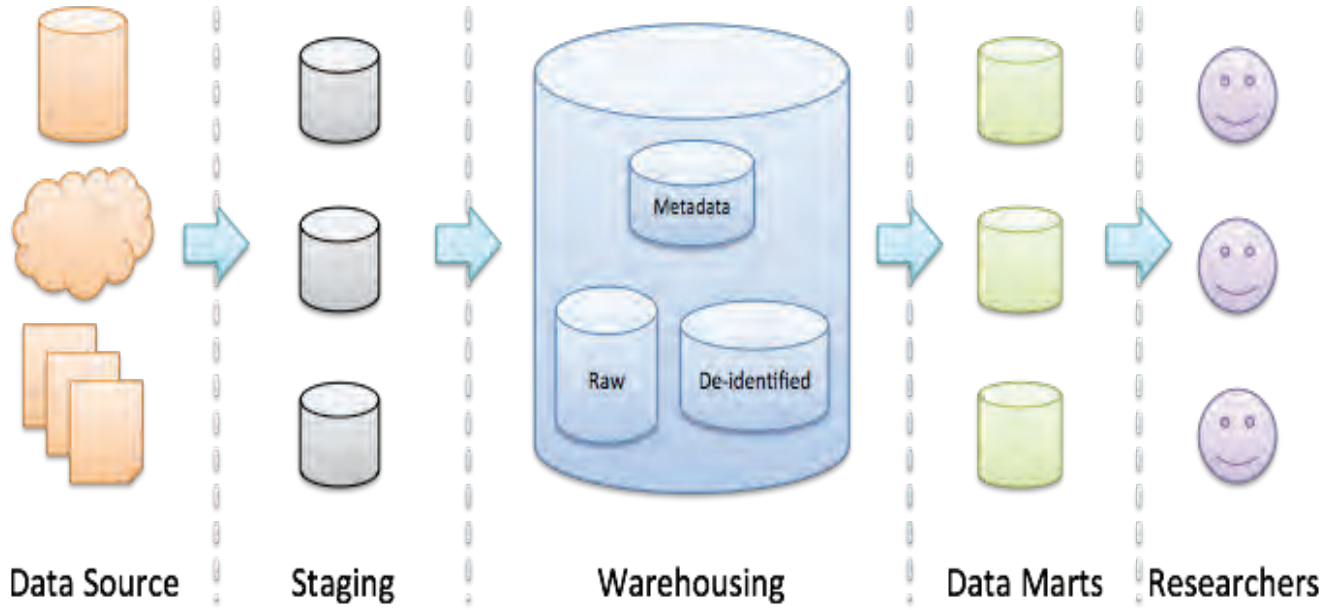


Figure 1: *CUSP Data Warehouse Facility Architecture*

Access Control

In addition to the protective structure of the data warehouse architecture, additional physical and electronic access controls will be necessary to ensure the security of the HUMAN project database. Given the value of the data and its highly sensitive content, it will be a very attractive target to malicious attackers, as well as "hacktivists" interested in embarrassing the research establishment to attempt disrupting the database to make a point, so it is critical that strong protections are put in place. Below are some of the approaches that we are considering implementing as part of our security plan.

We will, of course, implement the standard access controls for database security. Basic physical access controls used for sensitive data stores include securing the server rooms and limiting the number of people allowed to physically access the servers. Requiring double-identification methods for physical access to data storage systems (i.e. finger print and access card) is also likely. Physical access limits could also be imposed by air gap – a double firewall created by requiring researchers to use kiosks that access only the data marts, though this would likely only be practical when users wish to access the most sensitive data sets.

Electronic access should generally be granted on an as-needed basis, and broader access limited to a relatively small number of people who have passed strict background checks and regulated by role-based access privileges.

Double-identification systems will also likely need to be implemented for electronic access (i.e. fingerprint and password). For mission-critical functions, like those that could imperil the security of the entire database or the large-scale privacy of our subjects, "2-key" systems will likely be required. In such a system, only when two system operators who have both consented and been double-identified approve of an action, can a function of this category be implemented. (Sony's recent hacking by North Korea rested on its failure to use any 2-key systems at all).

However, even with careful administration of access privileges, it will be important to implement additional steps to protect the data from malicious insiders (or someone stealing the credentials of an insider's account, for lower security single-identification functions). Measures such as login monitoring and behavioral monitoring of administrators using automated systems that look for 'out of the ordinary' behavior can be very powerful and are standard in most high security data environments. Such systems will need to be implemented as well. It should be noted that these automated alerts for unusual patterns of behavior work best for long-term employees who can be easily and closely monitored, and for whom there are clear expectations of behavior. If we expect to allow scholars to access the database directly, it may be more difficult to use these approaches. For this reason (and others) it is unlikely that scholars will be allowed direct access to the actual database. Only when the highest possible level of security is operating, can critical privacy and security functions be accessible if we hope to maintain the degree of security necessary for this project.

Another potential source of protection might be a "gateway solution," where bandwidth control is used to restrict how much data can be moved off the server at one time. However, some experts were doubtful about the strength of this strategy, as recent security breaches have been executed by moving data in packets just a bit below the transfer threshold. Still, as an additional measure this could prove valuable.

Access and Use

Even the most perfectly secure data system, however, is only as secure as the uses to which it is put. The use of secure data marts and a state-of-the-art security environment can largely guarantee that only permitted uses of the data occur. But in order to assure the privacy and security of our subjects, it is essential that the study management govern the use of data effectively. To achieve that, there will need to be a committee and a process to evaluate who is granted access to specific classes of data and under what terms. This will be particularly important for defining the access privileges granted to visiting scholars and others who may initially be unknown to the study team – and thus will have limited physical and electronic access to our systems. Important criteria to consider include (but are not limited to) security clearance procedures for scholars and employees, credentialing, and background checks. Institutional review boards (IRBs) at researchers' home institutions could, in principle, provide some level of assistance with the initial vetting process, as it would not be possible to use the database for research without an IRB approved protocol. However, IRBs do not have sufficiently uniform standards to provide adequate information for making access decisions, so additional evaluation by a committee under the aegis of the study will still be necessary. A basic outline of the Credentialing Committee will be required for the CD1 document and a detailed overview of its policies and procedures will be required no later than the CD3 document.

Even after substantial vetting, external researchers, particularly those working off site, are likely to be the greatest vulnerability to the system. Training in security and data management prior to any data access will be critical to ensure that researchers comply with all policies and regulations. Once work with the data commences, activity of the researchers on the network will need to be closely tracked and access privileges re-evaluated every quarter.

An additional approach would be to encourage collaborations by external researchers with existing staff researchers. Such collaborations would provide additional protection by limiting the number of people who can access the data, but could also add value for researchers (particularly those who are inexperienced with handling large, sensitive data sets) by offering an opportunity to learn from an expert. This approach, while likely improving the quality of research using the database, requires a sufficient quantity of expert resources in-house to support a reasonable number of collaborative projects. We note that this model is similar to the work of "trusted telescope operators" in national astronomical observatories – specially trained engineers that facilitate the data gathering operations of large-scale telescope facilities. However, it is important to note that this approach requires the financial resources to support technical staff. (We note that this could be funded via data use fees of some kind, perhaps employing a sliding scale by which researchers with limited funds are subsidized by for-profit research entities.

Special Protections for Especially Sensitive Data

As is the standard protocol at CUSP and many other places, all data ingested into the database will be classified along the continuum from highly sensitive (and thus most stringently access restricted) to not at all sensitive (and thus publicly available). This provides the opportunity to increase security measures for the most sensitive data without creating excessive barriers to the use of less sensitive data. This may mean adding additional layers for separating sensitive data, triple de-identifying it with different codes for different types of data and using a 2-key system to access the most private sets of data. Separating out sensitive data with different levels of security would be necessary at the level of both data storage and data marts.

System Testing

Regular penetration testing of the security of the database is critical. Tests should include attempts at both physical and electronic attacks by highly skilled testers. It will be essential to develop a serious team of experts who periodically attempt to penetrate our systems. Quarterly reviews by this team are an essential part of any ongoing security plan.

Disaster Survivability-recoverability

Two general strategies are commonly used to promote data recovery following disaster: mirrored servers and back up to a separate physical medium that is subsequently stored offsite. We note that standard procedures break the datasets into a large number of encrypted components, each typically stored at different locations, which must be brought back together for reconstruction of the database. Such an approach is essential for securing off-site backups of any kind.

One traditional approach that employs this method has been to create a set of fragmentary data tapes, for example a set of 10, that are then stored in safe deposit boxes at multiple offsite locations. The encrypted fragmentary tapes are each useless alone. A minimum subset, for example any 7 of the 10, are required for database reconstruction with this method. One advantage of such an approach is that physical tapes are less accessible to hackers, since the stored data is not accessible online. However, that lack of accessibility also limits the availability of the back-up media if a restoration is required (particularly if tapes are stored in geographically distant locations). Physical tapes are also a weakness, as experts reported that despite the best intentions, they have a tendency to get lost. (This compromises recovery more than security for a properly encrypted system which requires that 7 tapes be brought together from different locations for reconstruction.) Costs for this strategy include the purchase, storage, and destruction of the physical tapes.

Back up via point-to-point transfer to remote servers requires that data be sent from the main server to a remote site with very high-level encryption and 2-factor authentication, often with a 2-key system required for recovery. Under this model, data is physically encrypted inside our secure facility and only leaves after that encryption process is complete. Keys for decryption are handled as physically secured objects requiring 2-key double authentication. This method has complementary strengths and weaknesses to physical backups. Because information is being transferred and stored online, it is vulnerable to hackers should they be able to defeat all of the security systems that are in place. However, the backup copy (or copies) cannot be physically lost, and it is more easily accessed if any restoration is required. In the case of the HUMAN project there may be some synergies with infrastructure currently being developed by CUSP and Internet2 that may argue for the remote encrypted server approach. It should also be noted that the Council expressed some consensus that this remote online server approach will likely be preferred. This issue will require further study and an initial plan will be required for the CD1 document. For either solution, it will be critical to complete detailed capacity planning to ensure that the chosen option is scalable to our needs in the CD2 document.

In either case, we will also require the development of a process for the creation of back-ups and restoration in case of disaster. The more complete (but consequently more expensive) choice is to have a hot site – this is a location where operations can be resumed after a disaster as soon as you can get people into seats at the new site. A less expensive option is a cold site, where the company that maintains the offsite back up is responsible only for restoring the data itself – the owner is responsible for providing a physical site where operations can resume. A cost-benefit analysis of all options will be necessary at the CD2 stage in order to determine the optimal strategy to ensure protection of the data from disaster and facilitate restoration of the data as smoothly and rapidly as possible.

Meeting Attendees

Hannah Bayer
Chief Scientist, Institute for the Interdisciplinary Study of Decision Making
New York University

Justin Brookman
Director, Consumer Privacy Project
Center for Democracy and Technology

Justin Cappos
Assistant Professor of Computer Science and Engineering
New York University

Miyoung Chun
Executive Vice President of Science Programs
The Kavli Foundation

Marti L. Dunne
Associate Vice Provost for Research Compliance and Administration
New York University

Paul W. Glimcher
Julius Silver Professor of Neural Science, Economics and Psychology
Director, Institute for the Interdisciplinary Study of Decision Making
New York University

Robert M. Goerge
Senior Research Fellow, Chapin Hall
University of Chicago

Lynn Goldstein
Chief Data Officer, Center for Urban Science + Progress
New York University

Thomas Hardjono
Technical Lead & Executive Director
The MIT Kerberos Consortium

Christopher Martin
Science Program Officer
The Kavli Foundation

Jules Polonetsky
Executive Director and Co-chair
Future of Privacy Forum

Sharif Taha
Science Program Officer
The Kavli Foundation

Hilary Wandall
Associate Vice President, Compliance and Chief Privacy Officer
Merck & Co., Inc.

Marcy Wilder
Director, Privacy and Information Management Practice
Hogan Lovells

Miriam H. Wugmeister
Chair, Global Privacy and Data Security Group
Morrison & Foerster

APPENDIX D

BLUETOOTH TECHNOLOGY REPORT

Introduction

A better understanding in human behavior can be used to improve our daily lives. We can apply what we learned about human health and behavior to inform and improve public policy. Imagine understanding human behavior in a deep way that enables scientifically driven public policy. Unfortunately, we still do not fully comprehend the roots of human behavior, how we make decisions and what forces shape those decisions.

With recent expansions in technology, the introduction of the Internet of Things (IoT), new methods for the management and analysis of big datasets, and advances in smartphones, we have the opportunity for large-scale information gathering at a relatively low cost. This project aims to better understand the complex, interwoven, causative roots of human health and behavior by gathering diverse information about a large number of families within a major metropolitan area at an unprecedented level of breadth and detail. The gathering will work as a comprehensive large-scale survey intended to characterize the major forces that shape the human condition. We will gather data from medical records, genetic and microbiome data, patterns of physical activity, psychological profiling, educational tracking, economic development, employment and social networks for people living in New York City. All the data will be combined to get unique insights into the factors that are known to influence individual health and behavior, but which have never been integrated on such a broad scale.

As part of the project, we will collect data of hundreds of families inside their homes to understand their behavior and the interaction between parents and their kids. To reach hundreds of families, our solution aims to be cheap, easy to deploy and autonomous. The architecture should not depend on WiFi, initial deployment setup or any initial training step. In addition, we want to individually track each person at room-level accuracy without previous knowledge of the house layout. Further, we will gather the position of the parents and their kids continuously, with granularity of at least once per minute. In this report, we study the feasibility of using indoor localization with current, existing technologies. Moreover, we performed a set of experiments in an office building to better understand the solution limitation and behavior.

This report is presented as follows: Initially, we list current existing related work on indoor positioning. Afterwards, we detail Bluetooth Low Energy (BLE) behavior and characteristics. We performed some experiments and detail the results in the following chapter. Moreover, we discuss the technology in general and propose a technology setup to be used in the project. Finally, we deliberate about privacy and security issues.

Related Work

Over the years, several techniques to achieve indoor localization have been proposed. Accurate indoor positioning and tracking may be still an open problem [COMP2014, COMP2015, EXP2015]; however, applications that require room-level, or even meter level, accuracy can make use of one of those techniques. Existing solutions use technologies that range from WiFi, geo-magnetic beacons, to sound signals.

Using a combination of radio frequency and ultrasound, Bodhi et al. [CRICKET2000, CRICKET2005] presented a solution that achieved indoor localization with error average lower than 10 cm. In addition, they presented a graph-based algorithm [CRICKET2005] that reduces the number of nodes deployed necessary for the localization. Tarzia et al. [SOUND2011] proposed a sound-based system that uses the acoustic background spectrum to fingerprint a room. This solution yielded 69% correct fingerprint matches with room-level localization accuracy. Prigge and How used a magnetic solution for a very precise local positioning [HOW2004]. Due to their distributed solution, the beacons need to maintain synchronization with each other making necessary the use of wiring such as cable or the building's electrical wiring. Pirkl and Lukowicz [MAGN2012] develop a resonant magnetic coupling consisting of a 16x16x16cm transmitter and 2x2x2cm receiver coil. Using these devices, they perform indoor positioning with error of less than $1m^2$.

Abrudan et al. [MAGLOC2015] uses magneto-inductive (MI) fields for 3D indoor localization, achieving positioning with error of 80cm. In their solution obstacles are not an issue, due to the MI characteristics static objects are largely seen as transparent. The transmitter consists of a 30cm wood box and the receiver is a small 12x7x2cm device with battery capacity up to 12 hours. Chung et al. [GEO2011] created a fingerprint map of the environment using the magnetic field affected by the building structure. To identify its current location, the user's device measures the magnetic signature of its surrounding and compare with all previously stored fingerprints. Experiments showed that this technique achieved an error within 1.64 meters 90% of the time.

Using Bluetooth technology, Bruno and Delmastro [BT2003] used a wired network of Bluetooth access point coordinated by a central machine for indoor localization with room-level accuracy. Further, Cho et al. [CHO2015] investigated the discovery probability and latency of Bluetooth Low Energy devices. After performing extensive experiments, one of their conclusions is that with increasing number of BLE devices, delays of device discovery show an exponential growth.

Adib et al. [ADIB2014, ADIB2015] uses the radio signal broadcast and its reflection to infer obstacles and their distance. The transmitter sends a narrowband signal whose carrier frequency changes linearly with time. Therefore, it compares the frequency of the signal transmitted with the signal received to get the time-of-flight. In addition, the solution identifies and removes most of the background interference caused by reflections on static objects and walls. Furthermore, they showed that their solution is not limited for indoor positioning applications [SMART2015]. The work shows how this technique can detect the heart rate and breathing of stationary people with median accuracy of 99%. However, the solution cannot distinguish detected people, making it hard to track them.

Some existing techniques leverage radio-frequency technology such as WiFi router and ZigBee to calculate the distance from a transmitter to a receiver. Bose and Foh [BOSE2007] proposed a simplified model based on Hata-Okumara model to calculate the distance between transmitter and receiver using the radio signal strength.

In contrast, Bahl and Padmanabhan [RADAR2000] proposed a model based on a Path-Loss Prediction model that compromises simplicity and accuracy. In their model, to take in account the effects of obstacles between the transmitter and receiver, they included a wall attenuation factor. Tests showed a median error resolution in the range of 2 to 3 meters. Faragher and Harle [FARAGHER2014] studied the impact of BLE devices on fingerprint-based indoor positioning. They performed a set of experiments testing the data sampling size, scan frequency and noise mitigation. Jianyong et al. [JIANYONG2014] experimented with different techniques to mitigate the noise. Initially, to take in consideration the environment noise, they calibrate the model for each device using a least square piecewise fitting function. They keep further improving its accuracy using active learning. Afterwards, the signal strength data is smoothed over time using a weighted distance filter. Results showed an accuracy error less than 1.5m.

Similarly, Anna et al. [ANNA2014] proposed a two phase solution to mitigate the problems with fingerprint-based indoor localization. During the offline phase it calibrates the model for each device. On the online phase they remove outliers and smooth the data over time. Halder et al. [HALDER2014] uses a similar solution, in addition to signal strength their model also takes in consideration the link quality indicator of each signal received. Both solutions [ANNA2014, HALDER2014] achieved average accuracy below 50cm.

Unfortunately, most of these solutions present characteristics that don't fit within our scenario. Since we plan to perform our study with thousands of families the solution needs to be easy to deploy. Therefore, we expect it to be infrastructure free, autonomous, cheap, and fast to setup. In addition, we expect that not all families will have WiFi networks in their homes, so we cannot rely on solutions that depend on it. Moreover, our solution shouldn't depend on training, specific device positioning or previously knowledge of the home layout. Further, the equipment needs to be small, wireless, and should work autonomously for a long period of time without maintenance. Finally, the technology needs to differentiate between people, track them, support large distances, and calculate indoor positioning at least at room-level. Based on these requirements and the only current solution that fits our needs is Bluetooth Low Energy.

Bluetooth Low Energy

Bluetooth Low Energy or Bluetooth Smart, is a wireless personal area network standard created in 2010 aimed for the Internet of Things. The main difference from BLE to traditional Bluetooth is that it has considerable reduced power consumption while maintaining a similar communication range. The Core Specification of Bluetooth mandates a minimum communication range of 30 feet [BT4_2015]. This new technology was created to be used in a range of situations, from proximity sensing in a store, to home automation (Figure 2).

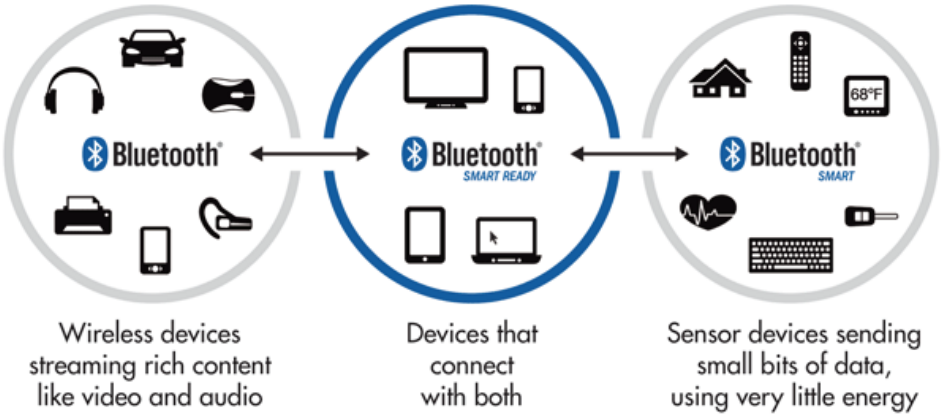


Figure 2: Bluetooth Smart compatible devices as part of a wireless ecosystem.

Bluetooth Smart devices communicate in two different ways. As broadcaster, the device keeps advertising information about itself and some data that can be consumed by the receiver (broadcaster in Figure 3). For example, clients inside a store may receive broadcasts containing information about items on sale. The smartphone, acting as observer, receives the data and decide which information is worth to consume or even discard.

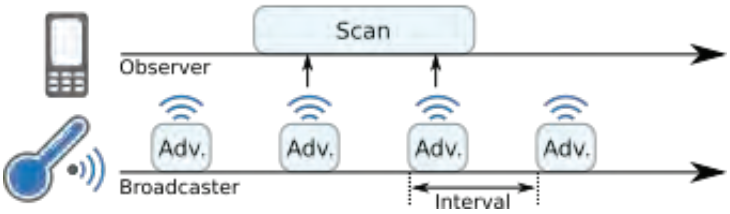


Figure 3: As a broadcaster beacon, the device keeps advertising data that may contain relevant information.

The smartphone can also establish a connection with an advertising BLE device (Figure 4). The connected device, acting as central, will have access to all the features and characteristics present in the beacon (acting as peripheral device).

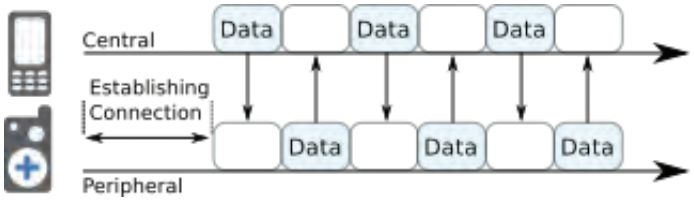


Figure 4: After establishing a connection, devices can exchange information by sending and polling data at regular intervals.

The central device can read and write any information to the beacon, collect all the data gathered by the sensors, control and order the peripheral to perform an action. A fitness tracker device is an example of a peripheral. The tracker will collect all sorts of data and will notify the cellphone whenever it has new data. The smartphone on the other hand, will request data whenever needed. It can also order the tracker to change its behavior or show a notification to the user.

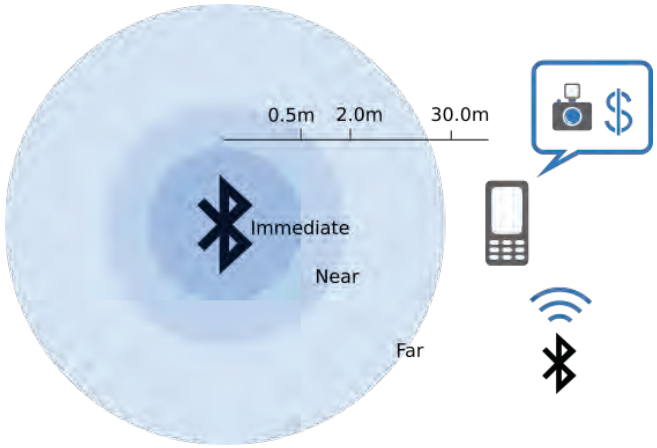


Figure 5: iBeacon proximity sensing divides the localization into three regions. Depending on the distance to the device it may trigger an action on the smartphone.

iBeacon is a proximity sensing protocol standardized by Apple. This protocol works over Bluetooth Low Energy and aims to trigger location-based action (Figure 5). For example, users can receive customized information based on its indoor localization. As an open-source alternative, AltBeacon [ALTBEACON15] was created aiming at not favoring any vendor over another. Likewise, Google created its own open-source standard called Eddystone [EDD15]. The difference between Eddystone and the other two is that it doesn't need any *app* installed to receive a notification. While in a Starbucks, for example, an iPhone user needs the Starbuck *app* to receive notification. On

the other hand, Eddystone sends to its customers an URL that opens in the browser, allowing notification to those that have not the *app* installed.

We decided to focus on the Bluetooth Smart core to have more freedom on the development, and to have more distance options (than just immediate, near and far), as shown as in Figure 5. Moreover, those standards add one more layer which is not necessary good for us. We want a solution that has the same behavior independently of the operating system and of the API being used.

Measuring Distance

Using the information of the Bluetooth Smart signal strength, we can use a radio frequency propagation model to calculate the distance between a transmitter (*tx*) and the receiver (*rx*). Existing solutions calculate the distance using different base models [BOSE2007, RADAR2000].

Bose and Foh [BOSE2007] proposed a formulation based on the Hata-Okumara model (Equation 1) to calculate the expected signal power at the receiver (P_{rx}). This model takes in consideration the log-based propagation model and uses the signal frequency information and the antennae gain (G_{tx} and G_{rx}).

(1) $P_{rx} = P_{tx} - 10 * n * log_{10}(d) + G_{tx} + G_{rx} + C$

Bahl and Padmanabhan [RADAR2000] simplified the Floor Attenuation Factor propagation model, taking in consideration the effects of indoor obstacles. In Equation 2, M and nW are the number of walls in the building, and W is the attenuation applied by each wall.

(2) $P(d) = P(d0) - 10 * n * \ log_{10}\left(\frac{d}{d0}\right) - \begin{cases} nW * W, nW < M \\ M * W, nW \geq M \end{cases}$

In both Equations 1 and 2, P_{rx} and $P(d)$ represent the power of the signal at the receiver, at the distance d. Both are log-based models for radio frequency propagation and can be generalized as the formula shown in Equation 3.

(3) $R = R_0 - 10 * n * log_{10}\frac{d}{d_0} + C$

The strength of the signal is directly affected by the Bluetooth hardware. For this reason, the model is calibrated for a specific hardware by measuring the signal strength (R_0) at a well-known location distance (d_0). To simplify the calculation, it commonly used the received signal strength indicator (RSSI) (R_0) measured at 1 meter ($d_0 = 1$). The n represents the attenuation factor and C, is a constant used to mitigate the effect of barriers, such as static objects and walls.

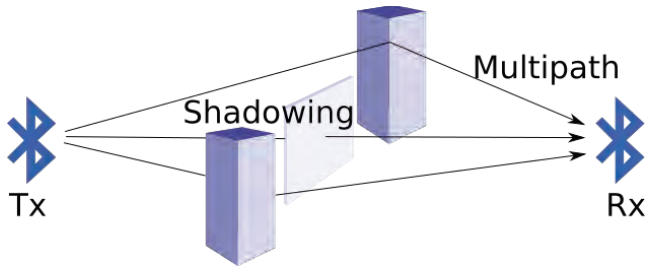


Figure 6: Signal path from transmitter to receiver.

Further, the strength of a received signal is directly affected by multipath propagation and shadowing (Figure 6). Multipath occurs due to the fact that radio signals are spread out of the transmitter in multiple directions. These signals may reflect of a variety of surfaces and reach the receiver via paths other than the direct light-of-sight. Shadowing is caused by *occluders* between the transmitter and the receiver that can attenuate or obstruct the radio signal.

In our scenario we want to deploy our system in an indoor environment, thus the number of occurrences of both cases increases. In the following sections, we will describe these problems; showing how they affect the data collected and how we can mitigate them.

Experiments and Results

In this section we detail and discuss the experiments we performed to better understand RSSI-based positioning and tracking. For this, we used StickNFind (SNF) beacons, a couple of fitness tracker devices (Fitbit Charge and Misfit Shine), a Samsung Galaxy S4 mini with Android 4.4.2 and an iPod Touch running iOS 8.3.

The experiments were performed in a slightly busy office building, containing different source of radio frequency interference that includes, not limited to, Bluetooth mouse and keyboards, WiFi routers and fitness trackers. In the following subsections, we detailed and discussed the experiments and results obtained.

Unique Identifier

In our solution, we expect to uniquely identify BLE-enabled devices in order to track them. Based on the Bluetooth 4.0 specification, beacons share their MAC address in every broadcast. This MAC address could be used as the unique ID to identify the device. Unfortunately, the address in the broadcast may be fake. For security reasons, the Bluetooth 4.0 specification states that hardware may opt to use a temporary address during the broadcast. If necessary, after pairing between two devices it is possible to request the correct MAC address.

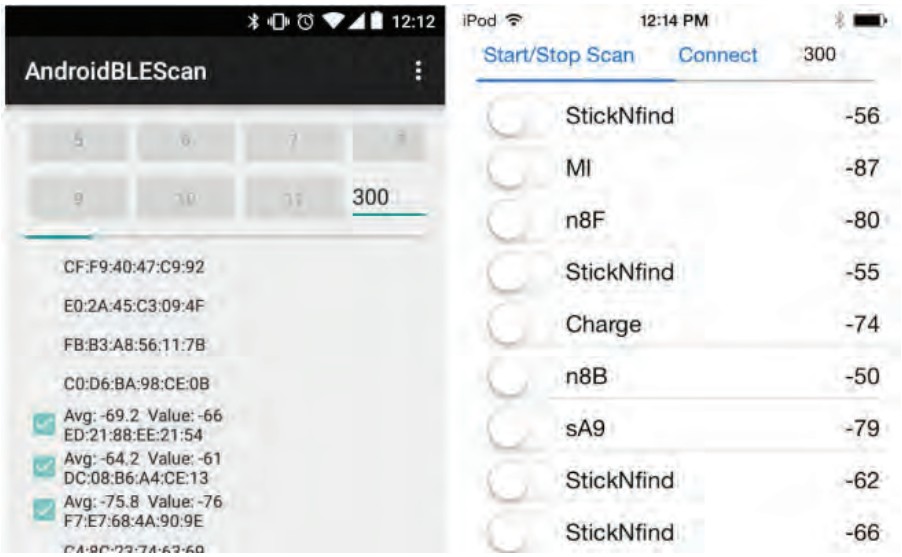


Figure 7: The left screenshot shows the *app* running on Android listing the MAC address of the beacons detected.

On the right, we show the *app* on iOS. Since we cannot access the MAC address on iOS, we are listing its not unique name.

During our tracking experiments, we were able to get the MAC addresses of all devices from the broadcasts using Android. In addition, neither of the tested beacons implemented the temporary MAC address solution. However, the same behavior was not detected during our experiments using the iPod Touch. Unfortunately, the iOS API do not expose the device address to the application layer. Instead, it uses an UUID (Universally Unique Identifiers) to uniquely identify each device. This UUID will be remembered the next time both devices come across each other. Since UUID are created with random bytes, different iOS devices will create distinct UUID for the same beacon.

Figure **7** shows the screen of the *app* running on both Android and iOS. As you may observe, since we can't get the MAC address on iOS we had to use the beacon's not unique name as a visual identifier.

Data Acquisition

Bluetooth Smart devices may be not connectable (e.g. beacons). For connectable devices, some operating system opts to keep only one advertisement packet during the discovery procedure. In addition, increasing the number of BLE devices also increase the discovery delay [CHO2015]. Therefore, instead of using discovery scan to get the signal strength, we opted to establish a connection between the smartphone and the beacon to constantly request the RSSI. Alternatively, we could restart the scan every second. In the future, we will perform a set of experiments to see which solution is more stable and consume less battery.

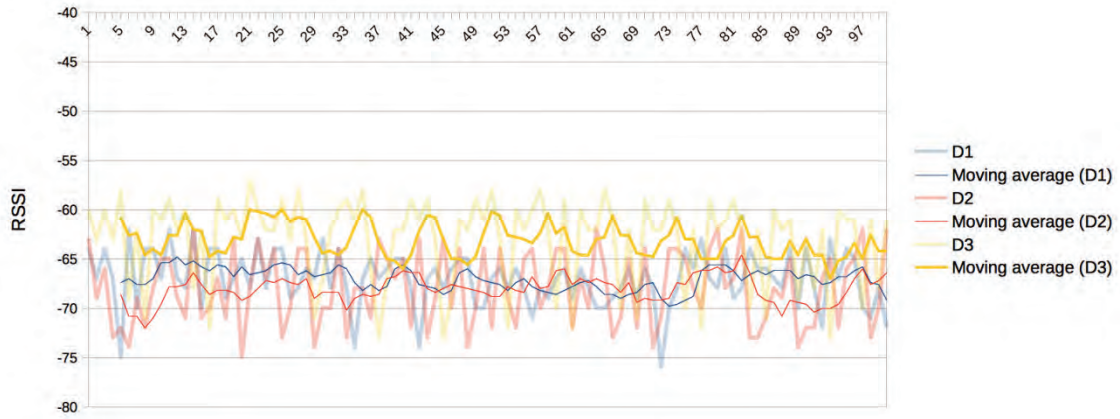


Figure 8: RSSI raw data collected showing multipath noise and the same data smoothed over time.

In this first experiment, we collected the signal strength of 3 SNF devices over time at the distance of 1 meter to identify how much multipath noise affects the data. Figure 8 shows the data collected and its moving average using 5 samples. In this figure, all peaks represent multi-path noise.

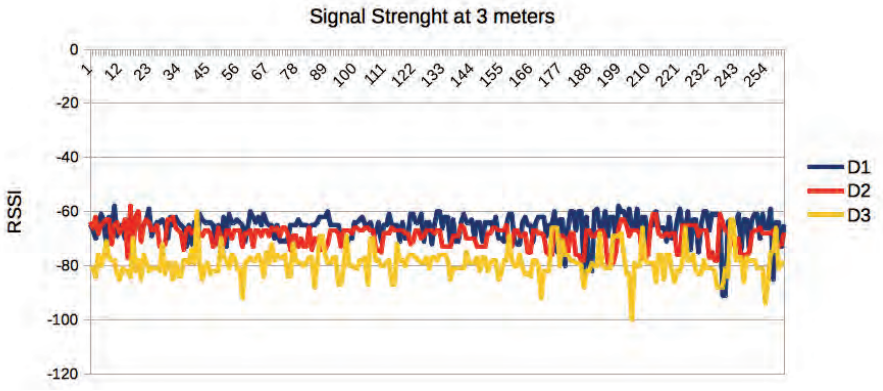


Figure 9: Signal strength captured using an iOS device at different distances.

In contrast to the results obtained using Android, RSSI collected using an iOS seemed more stable. As you may observe in Figure 9, the curve shows a small indication of noise. However, the noise on the data collected increases as the devices get further away. We believe that iOS may be applying some sort of a processing filter over the raw RSSI. A smoothing processing such as median filter would explain the unexpected *peaks* we see at the green curve in Figure 9.

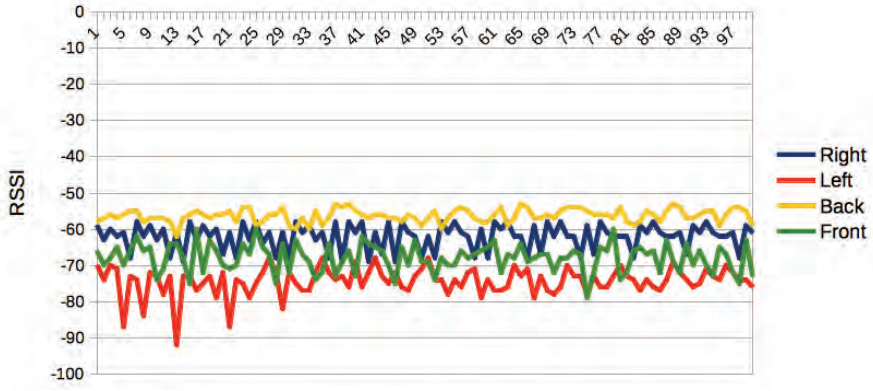


Figure 10: Average signal strength measured for the same device at different orientations.

During the first experiment we observed that, due to the location of the Bluetooth hardware in the smartphone, the device orientation affected the strength of the signal received. To further investigate this behavior, we performed experiments using 3 beacons, with 3 different orientations at 1m from the receiver. Results showed that there's a clearly gap in the RSSI measure. As we may observe in Figure 10, the difference in signal strength at close range may be as large as 10dBm. This is an expected behavior as showed by [DAHLGREN2014].

Interference

Obstacles like walls and furniture are common *occluders* that interfere with the radio signal propagation in an indoor environment.

Table 1 presents a merge of two reports [APPLE15, CUMBERLAND15] that list common barriers material, its potential to affect the signal, and an estimation of an attenuation value.

Table 1: Radio Frequency reflective and absorbing obstructions.

Type of Barrier	Potential	2.4 GHz Attenuation (db)
Interior Office Door	Low	4
Solid Wood Door	Low	6
Brick	-	6
Interior Office Window 1"	-	3
Glass Divider 0.5"	-	12
Interior Hollow Wall 4"	-	5
Interior Solid Wall 5"	-	14
Water	Medium	-
Marble	Medium	6
Metal	Very High	-

The following experiments investigate how the material of *occluders* affects the signal strength and the multipath noise. Initially, we placed a wooden furniture and a metal cabinet in between the transmitter and the receiver. Afterwards, we placed the beacons at 4 meters from the smartphone using a glass door and an office wall as barriers. As expected, we observed a reduction in the signal strength and an increase in the amount of multipath. As we may observe in Figure 11, even after smoothing the curve using moving average, the data is still strongly affected by the noise.

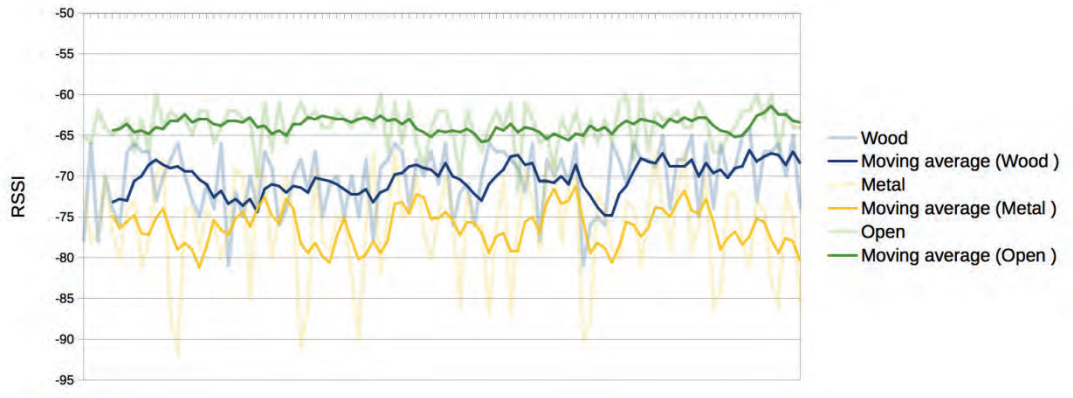


Figure 11: Signal strength affected by different barriers, composed of materials commonly found indoors.

Subsequently, we investigated how multiple beacons broadcasting at the same time affected the RSSI collected. We placed 3 devices side-by-side at 1 meter from the receiver exchanging data at 2Hz. The increase in the number of beacons broadcasting didn't affected the signal strength obtained and neither increased the number the multipath noise.

Finally, we noticed that experiments performed in different time of the day had different noise pollution. More specifically, we performed experiments at 3 different time periods, with 2 Bluetooth Smart devices side-by-side. The result showed that the amount of noise and the strength of the signal were directly affected by the changes in the environment, number of RF based devices around and the number of people in the office.

Delays

To improve the localization accuracy, we need to mitigate the existing problems affecting the signal strength. For example, to reduce multipath noise we can sample the measurements to calculate the RSSI, discard outliers and smooth the data over time. Thus, the number of samples per time affects the granularity of the data acquired. In the following experiments, we investigated the response delay during data acquisition and the time necessary to establish a connection.

We performed four experiments using distinct devices at 0.1m and 3 meters distance to the smartphone. Running on Android the average response time were in the range of 10 to 15ms, in all case scenarios (Table 2). In addition, to establish connection to 4 beacons at the same time, the smartphone needed 5.7 seconds. We performed the same experiment with the *app* running on the O.S. background and obtained a similar result.

Table 2: Delay in milliseconds to receive the signal strength indicator after requested.

	Distance of 0.1m		Distance of 3.0m	
	3 beacons	**6 beacons**	**3 beacons**	**6 beacons**
Average	14.6	10.82	14.8	10.87
Median	15	4	15	4
Maximum	32	56	35	73
Minimum	1	0	1	

Finally, the same experiments were performed using iOS. Unfortunately, the results obtained were different than expected. The delay to get the RSSI had a fixed interval. No matter how frequently requested, or how many simultaneous devices are connected, the delay to get the signal strength was always 1000ms (±15ms).

Calculating Distance

Radio frequency based indoor localization solutions commonly uses log-distance path loss model to calculate the distance from transmitter to receiver [RADAR2000, ANNA2012, DONG2012, JIANYONG2014, HALDER2014]. In this subsection, we experiment with this model and test its accuracy.

$$(4)\ R = R_0 - 10 * n * log10\left(\frac{d}{d0}\right) + C$$

$$log10\left(\frac{d}{1}\right) = \frac{R - R_0 - C}{-10 * n}$$

$$d = 10^{\frac{R - R_0 - C}{-10 * n}}$$

Due to hardware differences, for each device we need to calibrate the model using RSSI at 1 meter distance (Measuring R_0 using d0 = 1, Equation 4). In this first experiment, we measured the signal strength of two fitness tracker, with different hardware, at a fixed distance. The results showed that the difference in strength between hardware can be as large as 20dBm (Figure 12).

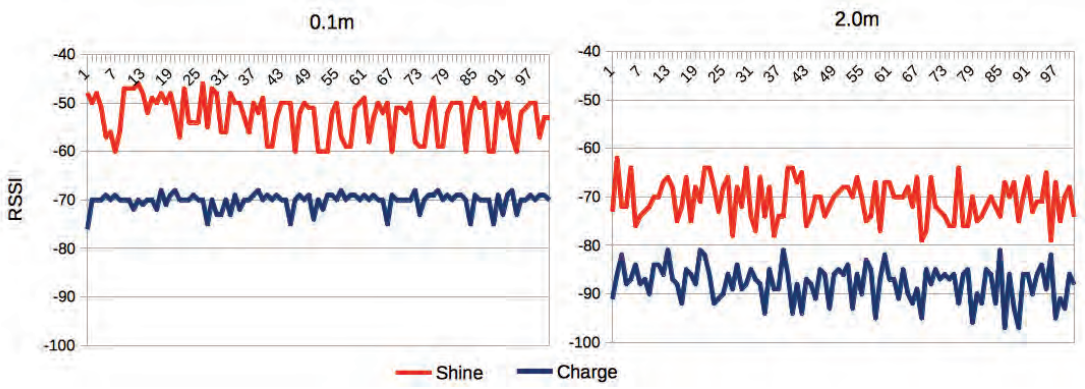


Figure 12: Signal strenght of a Misfit Shine and Fitbit Charge measured at 0.1 and 2m.

Afterwards, we collected the signal strength of 6 beacons at 11 line-of-sight distances. This experiment tests the accuracy of the log-distance path loss model (Equation 4). In Figure 13, each curve represents the RSSI at each location, for each device. We can observe the log-shaped curve behavior showed by each curve.

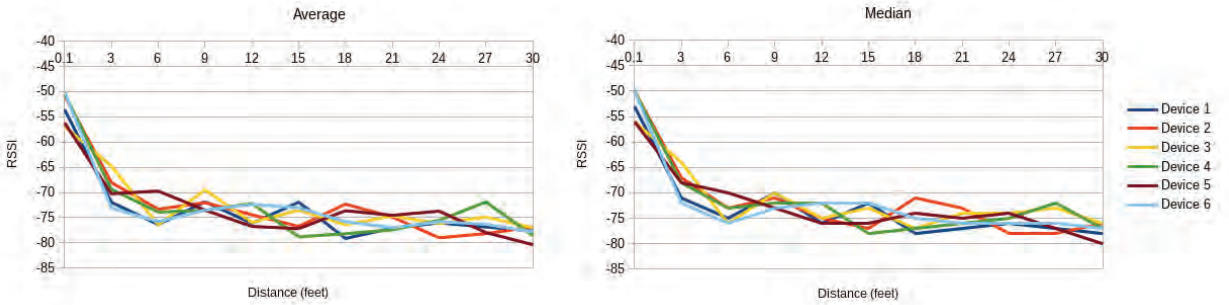


Figure 13: Average and Median value of the signal strength acquired at different distances.

Multipath and shadowing are well known problems to RSSI-based localization and different approaches have been proposed to mitigate them [RADAR2000, FARAGHER2014, JIANYONG2014, ANNA2014]. However, for the following experiment we decided to verify the precision of the calculated distance without any of those methods. For this reason, we are using a fixed value for the attenuation factor ($n = 2.5$), and a constant value for the modifier ($C = 0$).

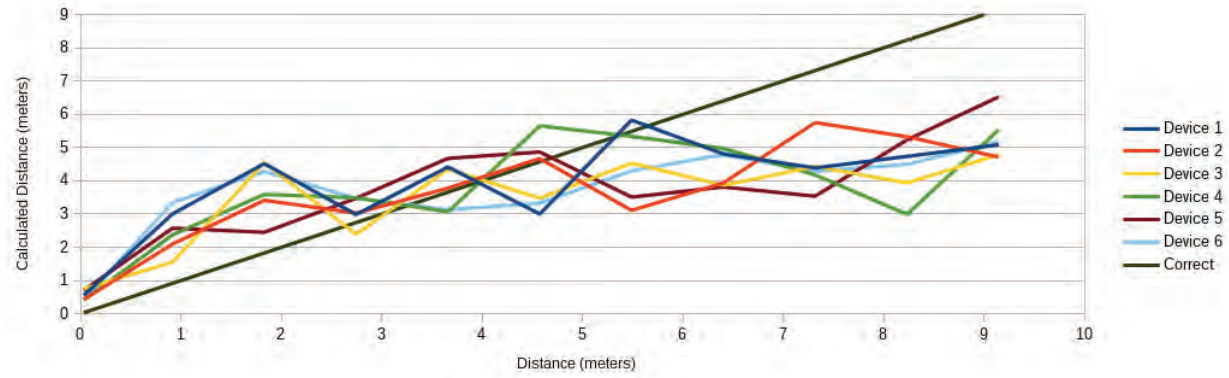


Figure 14: Comparison between calculated and actual distance using 6 devices.

As we may observe from Figure 14, the calculated distance are often overestimated at closer distance and, after around 5 meter, the model underestimate it. Even though we are using a generalized version of the log-model propagation, we obtained a behavior similar to the reported by Bose et al. [BOSE2007]. To mitigate this behavior, Bose and Foh proposed a dual model approach that adapts *n* according to the signal strength.

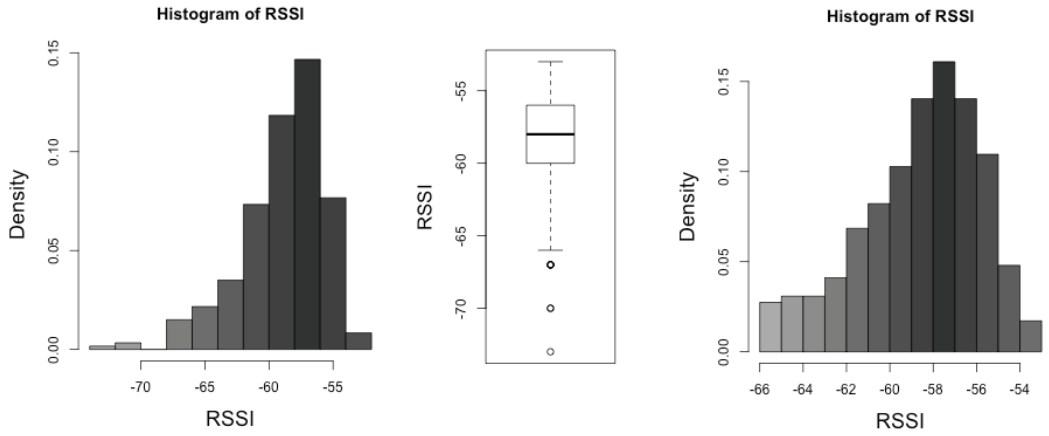


Figure 15: On the left we show the histogram of the data collected and, on the right, we have the histogram after the identification and removal of outliers.

In this experiment, without applying any technique to handle the noise, we obtained an average error of 1.77m and median of 1.56m. As showed by Zhu et al. [JIANYONG2014], a Gaussian distribution can reflect the randomness of the RSSI obtained. Thus, we believe that our result can be improved by identifying and removing outliers from the sample data we collected (Figure 15)

Information Generation

Based on the results obtained in our experiments, we recognize the potential of existing beacon devices as data gathering/generator in our scenario. Therefore, we envision the generation of the following information using Bluetooth Smart devices: a layout graph representing the indoor environment; the smartphone positioning in the graph; a list of moving devices with their current distances, as well as their moving directions; the current instance including the distance to all devices.

In this section, we deliberate on the data generation and how the created content can be used. We describe the use of multiple devices to generate a layout representation of the smartphone surroundings. In addition, we describe the process to distinguish between stationary and moving devices and how this classification can be used later on to help identify devices on people. Furthermore, we detail some of existing open sensors that may improve the data we plan to gather in our scenario.

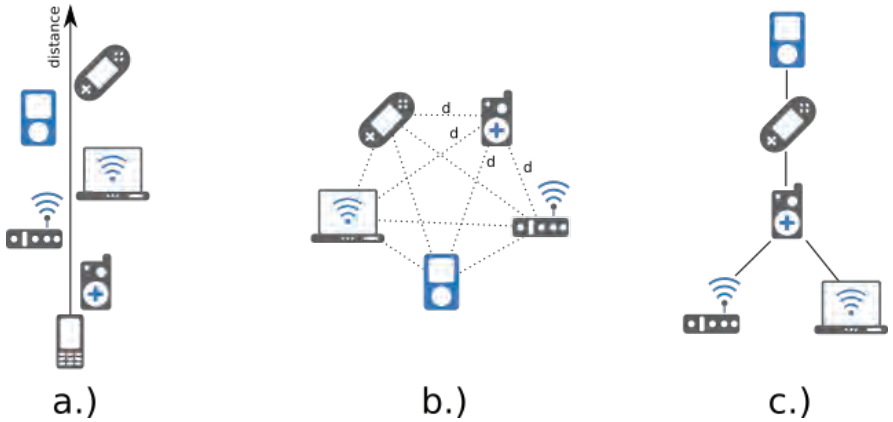


Figure 16: At first we calculate the distance from the smartphone to all devices. Then, we measure the distance between all of them. Finally, we use historical data and build a layout graph.

Distance to Device

Initially, we calculate the distance of every device to the smartphone (Figure 16a). The distance of existing beacons to the mobile will be stored, since that the historical data is important to extract more information from the scenario.

Distance Between Devices

To determine the distance between all devices we use the historical data (Figure 16b). We suppose that at one point in time, the smartphone will be immediately close to a device. At that point, the device distance to all other BLE will be the mobile phone distance to them. The distance between all beacons is a matrix that fingerprints the environment configuration.

Stationary Devices

To identify if a BLE-enable device is stationary or not, we can use the processed distance grid, historical data and the current beacons' distance.

Initially, while we do not have the fingerprint matrix or it is not up-to-date (e.g. we discovered a new device), this device is classified as non-stationary. Afterwards, if the mobile phone is not in movement, any device changing its distance by a certain value that exceeds a threshold will also be considered as non-stationary. Moreover, with the distance matrix measured, the next time the smartphone gets close to a device, we compare the current instance with the previously calculated fingerprint matrix. Any inconsistent value will indicate that a device is non-stationary.

This status distinction enables moving devices to be handled different. For example, stationary objects may be daily IoT devices around the house. For moving objects, on the other hand, there's a high chance that the device is being carried by a person.

Environment Layout

Following, we generate the graph representation of the environment (Figure 16c) where each node is a device and the link between then are *physical connections*. We assume that two devices are *physically accessible* if, at some point in the historical data, the smartphone was able to go from one device to another without getting too close to any other beacon (Figure 17).

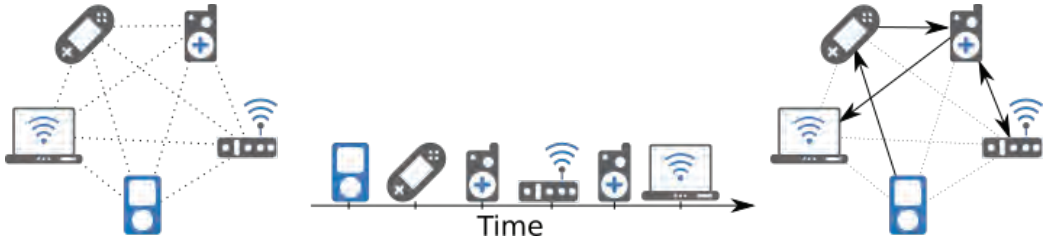


Figure 17: To determine connectivity status between nodes, we use the historical data and check the smartphone flow along the time.

Generated Information

We propose the generation of the following content: a layout graph containing the node connectivity and their distance; the current smartphone position within the graph; the list of probable locations of all others mobile BLE-enabled devices in the layout graph; the current instance containing the distance from the mobile to all beacons; and the list of dynamic devices with their current distances and moving directions.

Figure 18: We can search discovered BLE devices for well-known characteristics, by using their sensors without knowing the device or its API.

Others

In addition to the sensors within the smartphone (e.g. accelerometer, gyroscope, magnetometer, light sensors), we can search and use sensors available in IoT devices around us (Figure 18). We do not need to have previously knowledge of the device or use its specific API. There are some characteristics (sensors) commonly adopted for Bluetooth devices [BLE_CHARACTERISTICS]. In this list of characteristics we can acquire information, such as: weight, temperature, rain fall, humidity, etc.

Discussion

In this work, we tested the viability of using Bluetooth Smart devices in our study scenario. The results obtained are encouraging; we were able to calculate the distance between BLE devices with accuracy median of 1.56 meters. After analyzing the results, we can imagine the following setup to collect the data we need.

Initially, we can uniquely identify each room of the house by placing a beacon-like BLE device inside it. Based on the distance of the smartphone to each beacon, we can discover in which room the device is. To be tracked, each occupant of the house would wear a fitness tracker. In addition to its distance to the smartphone, our *app* would also collect other information that may be available by the tracker (e.g. pedometer, heart-rate measure, temperature).

In our results, we noticed that the maximum average time delay to get the RSSI is 15ms. We can calculate how fine the data granularity can be based on the average delay and the number of samples we need for each reading. For example, if we use 10 samples to calculate the distance, we can acquire a distance reading each 1.5 seconds. However, what would directly influence the granularity of the data is the battery usage by the Bluetooth hardware and other sensors, and the amount of data that needs to be collected, processed and stored.

Theoretically, there's no limit to the number of simultaneous connection a single device can have [BT4_2015]. Although, in our experiments we noticed that Android O.S. hardcoded the maximum number of connection. The Android device used in our experiments had a limit of 6 simultaneous connections. Since the delay to get the signal strength is low and, depending of the granularity of the data we want, we can work around this issue by alternating connection between devices. Furthermore, we can create a priority queue, sorting the beacons using any heuristic that we see fit.

As expected, obstacles and barrier reduced the strength of the signal. Since we will only see this phenomenon for beacons that aren't in the same room as the smartphone, the presence of walls will improve the room detection. Although, it also affects the accuracy of the distance calculated from the smartphone to anyone in a different room.

With this devised setup, we will be able to know the cellphone location in the house. However, the parents may not be carrying it all the time while at home. In the worst case, both parents' cellphones may be stationed in the same room. Fortunately, since the parents also wear a fitness tracker, we may be able to infer their location using the graph-based solution.

To get the parents location, we build the connectivity graph representing the house using the data previously collected. Afterwards, knowing where the smartphone is located and its distance to the parents, we can use this graph to get a small list of possible rooms where they may be. The number of possibilities will further reduce if we compare this list with a list computed by different smartphone devices in different room. Moreover, we can use spatial coherence to track and infer the parents' location.

Some of upcoming decisions regarding our solution may be a compromise between battery consumption and data collection. Unfortunately, with our current configuration using Android 4.4, it's not possible to get a precise measurement of the battery. In the future, we will look for alternatives to measure the battery consumption. In addition, we plan to improve the accuracy of the distance calculated by taking in consideration obstacles [RADAR2000, ANNA2014]. Moreover, we can reduce the background noise using filtering algorithms (e.g. moving average, weighted-average filter, etc.). Finally, as showed by [JIANYONG2014], a Gaussian distribution

can reflect the randomness of the RSSI obtained. Therefore, we can identify and remove outliers from the sample to mitigate multi-path [ANNA2014, JIANYONG2014].

Privacy and Security

Due to the Bluetooth Smart characteristics, some potential privacy issues arise. Since all the information and communication happens in a wireless network, it's possible to intentionally listen to a communication between two devices. In addition, it is possible to impersonate and act as someone else. For example, it would be possible to not only intercept the data the fitness tracker is sending to the smartphone, but also impersonate the cellphone and order the tracker to perform actions. Moreover, since the device MAC address is unique, it's possible to use it to track people.

Fortunately, those problems have been addressed by the Bluetooth specification. To increase the security BLE devices can use encrypted communication. After pairing and exchanging public/private keys, connected devices will communicate using an AES-128bit encryption. In addition, BLE devices are not required to broadcast its true unique MAC address (this is optional after Bluetooth 4.2 specification). Instead, it creates a temporary MAC address. It is still possible to get the device unique MAC address later on during pairing procedure. Moreover, modern smartphones with Bluetooth Smart capability allow the user to toggle the Bluetooth visibility to avoid tracking.

References

[ADIB2014] ADIB, Fadel, KABELAC, Zachary, KATABI, Dina, MILLER, Robert C. "3D Tracking via Body Radio Reflections". *USENIX Symposium on Networked Systems Design and Implementation*, NSDI'14, 2014.

[ADIB2015] ADIB, Fadel, KABELAC, Zachary, KATABI, Dina. "Multi-Person Localization via RF Body Reflections". *USENIX Symposium on Networked Systems Design and Implementation*, NSDI'15, 2015.

[ALTBEACON15] http://altbeacon.github.io/android-beacon-library/

[ANNA2014] HEINEMANN, Anna, GAVRIILIDIS, Alexandros, SABLIK, Thomas, STAHLSCHMIDT, Carsten, VELTEN, Jörg, KUMMERT, Anton. "RSSI-Based Real-Time Indoor Positioning Using ZigBee Technology for Security Applications". *Multimedia Communications, Services and Security Communications in Computer and Information Science*. 2014. pp 83-95. Springer Int.

[APPLE15] - "Potential Sources of Interference". Apple Inc. https://support.apple.com/en-us/HT201542

[BOSE2007] BOSE, Atreyi, FOH, Chuan Heng. "A Practical Path Loss Model For Indoor WiFi Positioning Enhancement". *International Conference on Information, Communications and Signal Processing*, 2007, IEEE.

[BATTERY15] https://source.android.com/devices/tech/power/index.html#device-power

[BLE_CHARACTERISTICS] https://developer.bluetooth.org/gatt/characteristics/Pages/CharacteristicsHome.aspx

[BT2003] BRUNO, Raffaele, DELMASTRO, Franca. "Design and Analysis of a Bluetooth-Based Indoor Localization System". *Personal Wireless Communications. Lec*. in Computer Science Vol. 2775, 2003, pp 711-725.

[BT4_2015] https://www.bluetooth.org/en-us/specification/adopted-specifications

[CHO2015] CHO, Keuchul, PARK, Woojin, HONG, Moonki, PARK, Gisu, CHO, Wooseong, SEO, Jihoon, and HAN, Kijun. "Analysis of Latency Performance of Bluetooth Low Energy (BLE) Networks". *IEEE Sensors (Basel).* 2015. 15(1): 59-78.

[COMP2014] Microsoft Indoor Localization Competition. http://research.microsoft.com/en-us/events/ipsn2014indoorlocalizatinocompetition/

[COMP2015] Microsoft Indoor Localization Competition. http://research.microsoft.com/en-us/events/indoorloccompetition2015/

[CRICKET2000] PRIYANTHA, Nissanka Bodhi, CHAKRABORTY, Anit, BALAKRISHNAN, Hari. "The Cricket Location-Support system", *6th ACM International Conference on Mobile Computing and Networking*, Boston, MA, August 2000.

[CRICKET2005] PRIYANTHA, Nissanka Bodhi, BALAKRISHNAN, Hari, DEMAINE, Erik, TELLER, Seth. "Mobile-Assisted Localization in Wireless Sensor Networks", *Proceedings of IEEE Conference on Computer Communications, IEEE INFOCOM 2015*, March 2005.

[CUMBERLAND15] - "Report: How Signal is Affected". www.ci.cumberland.md.us

[DONG2012] DONG, Qian, DARGIE, Waltenegus. "Evaluation of the reliability of RSSI for indoor localization". *International Conference on Wireless Communications in Unusual and Confined Areas (ICWCUCA)*, 2012. pp. 1-6.

[EDD15] https://developers.google.com/beacons/

[EXP2015] LYMBEROPOULOS, Dimitrios, LIU, Jie, YANG, Xue, CHOUDHURY, Romit Roy, HANDZISKI, Vlado, SEN, Souvik. "A Realistic Evaluation and Comparison of Indoor Location Technologies: Experiences and Lessons Learned". *The 14th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN'15)*. ACM 2015.

[FARAGHER2014] FARAGHER, Ramsey, HARLE, Robert K., "An Analysis of the Accuracy of Bluetooth Low Energy for Indoor Positioning Applications", *Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2014)*, Tampa, Florida, September 2014, pp. 201-210.

[GEO2011] CHUNG, Jaewoo, DONAHOE, Matt, SCHMANDT, Chris, KIM, Ig-Jae, RAZAVAI, Pedram, MICAELA, Wiseman. "Indoor Location Sensing Using Geo-magnetism". *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. MobiSys'11*. 2011. pp 141--154.

[GOOGLEPLAY15] - https://developer.android.com/about/dashboards/index.html

[HALDER2014] HALDER, Sharly Joana, GIRI, Paritosh, KIM, Wooju. "Advanced Smoothing Approach of RSSI and LQI for Indoor Localization System". *International Journal of Distributed Sensor Networks,* 2014.

[HOW2004] PRIGGE, Eric A., HOW, Jonathan P. "Signal Architecture for a Distributed Magnetic Local Positioning System". *IEEE Sensors Journal*, Vol. 4, No. 6, pp 864-873. December 2004.

[JIANYONG2014] ZHU, JianYong, HAIYONG, Luo, CHEN, ZIli, LI, Zhaohui. "RSSI Based Bluetooth Low Energy Indoor Positioning". *5th International Conference on Indoor Positioning and Indoor Navigation. IPIN 2014*, 2014.

[MAGLOC2015] ABRUDAN, Traian E., XIAO, Zhuoling, MARKHAM, Andrew, TRIGONI, Niki. "Distortion Rejecting Magneto-Inductive 3-D Localization (MagLoc)". *IEEE Journal on Selected Areas in Communications*, Issue 99. 2015.

[MAGN2012] PIRKL, Gerald, LUKOWICZ, Paul. "Robust, low cost indoor positioning using magnetic resonant coupling". UbiComp 2012. pp 431-440.
[RADAR2000] BAHL, Paramvir, PADMANABHAN, Venkata N. "RADAR: an in-building RF-based user location and tracking system". Institute of Electrical and Electronics Engineers, Inc. 2000.
[SOUND2011] TARZIA, Stephen P., DINDA, Peter A., DICK, Robert P., MEMIK, Gokhan. "Indoor Localization Without Infrastructure Using the Acoustic Background Spectrum". Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. MobiSys'11. pp 155-168. 2011.
[SMART2015] ADIB, Fadel, MAO, Hongzi, KABELAC, Zachary, KATABI, Dina, MILLER, Robert C. "Smart Homes That Monitor Breathing and Heart Rate". ACM CHI'15, 2015.

APPENDIX E

WEARABLE TECHNOLOGY REPORT

July 8, 2015

Introduction

In the early stages of the Kavli HUMAN Project design process, wearable activity trackers were proposed as a measure for activity levels and sleep behavior. During discussions of the Measurement and Technology Advisory Council meeting there was some speculation amongst the experts at the meeting that the feasibility of administering these measurements longitudinally across the Kavli HUMAN Project population would be low. It was noted by those at the meeting that wearable devices often have unsuccessful long-term compliance rates, therefore making the probability of obtaining comprehensive data for longer than a few weeks at a time unlikely.

In order to determine what factors were most influential in compliance for wearables, to discover potential problem points in their use, and to inform future participant incentive strategies for increasing compliance, we began a pilot test of 10 kinds of wearable devices, with a non-random sample group of 22 individuals (and 1 pet.) The pilot ran approximately 8 months, beginning on November 14, 2014.

Technology Selection

Wearable technology most frequently offers users a look into their activity levels and sleep patterns, along with various additional features unique to each wearable device. The most basic components are an accelerometer – which provides a continuous signal that gets converted to a discrete step count before the data comes off the band, and a Bluetooth beacon for sending the data to a phone or computer. Some wearables utilize physical measurements, such as heart rate or skin temperature and perspiration, to increase accuracy of their data. Certain wearables allow users to set activity or sleep goals, which it then monitors the progress of. Other features that may be included are watch capabilities, alarms, caller ID or text alerts, email notifications, and calendar access.

Ten kinds of wearables were selected based on their popularity and useful features. Wearable technology is on the rise, and a sample of the most popular and related devices was decided upon to be tested to determine their usability for the Kavli HUMAN Project. Wearable preference is a highly subjective matter, with different people prioritizing different measurements and factors of the devices. With this in mind, there is no wearable that can suit every person's subjective preference for personal use and therefore we did not seek to reach a consensus of the "best" device. This pilot of ten wearables sought to find a broadly workable solution in terms of collecting data for the Kavli HUMAN Project.

Basis Peak (2 devices) measures steps, heart rate, calorie burn, perspiration, skin temperature, and sleep duration and quality, while maintaining personally set activity goals. It also functions similarly to a phone, managing text messages, phone calls, emails, and calendar events. The device has a battery life of up to 4 days and is water resistant to 5 ATM. It comes with a silicone strap and a magnetic charger.

Fitbit Charge (3 devices) tracks calorie burn, automatically monitors sleep, maintains workout intensity, maximizes training, and optimizes health with charts and graphs. The device also functions as a caller ID, watch, and alarm. Goals can be set alongside measurements and manually recorded workouts. It is made of a water resistant (up to 1 ATM), flexible, durable elastomer material with a surgical-grade stainless steel buckle and a battery life of 7-10 days. The *Fitbit Charge HR* (1 device) is nearly identical, but uses automatic and continuous heart rate monitoring to improve accuracy of measurements. The device's battery lasts up to 5 days at a time.

Mi Band (3 devices) by Xiaomi Tech monitors daily activity levels, distance, and calories burned in collaboration with personal set goals. It automatically monitors sleep duration and cycles to inform app-based plans to improve sleep quality. It is equipped with a dual-process alarm, with both a pre-vibration and a sound alert. The device also vibrates with incoming calls. The band is water-resistant, made of hypoallergenic silicone and an aluminum alloy core disk with a 90 day battery life.

Microsoft Band (2 devices) by Microsoft Health tracks heart rate, steps, calories burned, and sleep quality. The band also provides a watch, timer, alarm, email previews, texts or calls, calendar alerts, and the weather. The device has a specific function for tracking golf, mapping traveled routes, and personal fitness workout tips. It is made of thermal plastic elastomer, dust and splash resistant, with a battery life of 48 hours of normal use and a charge time of 1.5 hours.

Misfit Shine (3 devices) by Misfit Wearables is advertised as a waterproof (50 m) activity tracker measuring steps, calories burned, distance, and sleep quality and duration. The app allows the wearer to set goals and monitor the progress of those goals, as well as to specify moments of walking, running, swimming, and cycling. The disk has a watch function and the battery can last up to 6 months, with a standard watch battery needing to be replaced when it dies. Made of Anodized Aircraft-grade Aluminum, the Shine is lightweight at 9.4 grams and able to be worn around the wrist or clipped to clothing. The *Misfit Flash* (1 device) varies as it is made with TPU/Polycarbonate, weighing 6 grams and has a lesser water tolerance (30 m).

Narrative Clip (2 devices) is a small, automatic 5-megapixel camera and app that video records the world around you. The device weighs 20 grams, has a 2-day battery life, and has a storage capacity of 8000 pictures. The camera automatically takes two pictures every minute, and can be manually activated or deactivated at will. The pictures are transferred and the device is charged by connecting to a computer via a USB cord. Pictures can be viewed on the computer or on a cell phone using the Narrative app. A newer version of the Narrative Clip can wirelessly transfer pictures to a cell phone.

StickNFind (6 devices) is a Bluetooth beacon that can be clipped onto a key chain, worn as a necklace, or attached directly to a device. Each device weighs 4.5 grams and is roughly the size of a quarter. An app allows you to track the location of the device with a live on-screen locator (with a range of approximately 100 feet), as well as page the device with vibrations or lights. The app also has the ability to "leash" with the device, so if the user moves

out of range they are notified with a custom alarm. The battery is a standard watch battery and can last for up to one year.

UP24 (3 devices) by Jawbone tracks activity and sleep, with the added benefits of alerts or alarms and information how to obtain goals and nutrition. The app allows the user to record duration and effort levels of each workout. The device is made of a hypoallergenic TPU rubber band with a TR-90 Nylon and nickel cap. It is water resistant, though unable to be submerged. The battery is meant to last 14 days, with a charge time of 80 minutes.

Method

The devices were distributed across the group based on interest levels. Pilot participants were instructed that they would be providing feedback, but were free to cease use whenever they chose to do so. This allowed participants to terminate their use for their own reasons, giving us a range of possible issues we could encounter in study participants. After they were finished with one device, they had the option to try another. Feedback was collected in both structured (surveys) and unstructured (free comment) forms. The structured surveys were sent out once a week, covering the topics: ease of set-up, use, comfort, battery capacity, drain on smartphone, influence of device on activity, and open-ended opinions on the devices. Unstructured feedback was welcomed via email, particularly from participants having tried multiple devices and wishing to compare them. Upon completion of a device, the participant was asked for the reason they stopped using it and for any further comments on their experience with the wearable.

Fig. 1

Photo courtesy of mybasis.com

Results

For our purposes, the true strength of the results was in the qualitative reporting on each device. In many cases, experience with a particular device varied across individuals to some degree.

Figure 1: *Basis Peak* (4 users) was one of the most interesting devices to wear, however wearers found the app difficult to use. The information gathered was also viewed as uninformative, poorly presented, and inaccurate. The device itself was not comfortable and too bulky to wear daily. Wearing ended when someone else wished to try the device, it broke, or became difficult to use and sync.

Fig. 1 Averages: Set-up 4, Usability 3.8, Comfort(day) 2, Comfort(night) 2.5, Application 1.7, Rating 5, Battery 3.5 days, Worn 7.1 weeks

Figure 2: *Fitbit Charge* (7 users) was found to be user-friendly, fairly easily paired with a small lag to sync. It was mostly comfortable, though some wished for it to be thinner and waterproof. Users liked seeing their step count,

Fig. 2

Photo courtesy of fitbit.com

but found the data to be coarse and inaccurate. The app was particularly well-liked and easy to navigate. These were worn until the users either lost interest or started another wearable.

Fig. 2 Averages: Set-up 2, Usability 1.6, Comfort(day) 3.2, Comfort(night) 3.8, Application 3.9, Rating 3, Battery 7 days, Worn 9.6 weeks

Fig. 3

Photo courtesy of fitbit.com

Figure 3: *Fitbit Charge HR* (3 users) was found to be fairly easy to set up and use, though clearing old data from a previous user was impossible. Users found it thick, getting caught on things, and though at the start it was unobtrusive and comfortable, it became uncomfortable to wear. It was noted that taking it off to shower or do dishes was a major drawback. Better data visualization and analysis was wished for, but the data over long-term was noted to be good. The wearers stopped using the device to try another one.

Fig. 3 Averages: Set-up 1, Usability 1.3, Comfort(day) 3, Comfort(night) 3, Application 3.8, Rating 3.3, Battery 5.2 days, Worn 8.7 weeks

Fig. 4

Photo courtesy of mi.com

Figure 4: *Mi Band* (3 users) was simple to set up and has simple, minimal information. Users found it uncomfortable for sleep and typing on a computer for extended periods of time. The disk that slides into the band was not secure, and one device was lost while the others fell out repeatedly. The only user who ceased wearing it lost the disk.

Fig. 4 Averages: Set-up 4.7, Usability 5, Comfort(day) 2.5, Comfort(night) 3, Application 3.5, Rating 4, Battery 7 days, Worn 5.3 weeks

Fig. 5

Photo courtesy of microsoft.com

Figure 5: *Microsoft Band* (2 users) was easy to set up, but took a large amount of time to sync. The ability to customize the device and the plot of activity and heart rate overlapping were particularly liked. The band was uncomfortable and difficult to loosen once on, and the face of the device was too fragile. Users would have liked it to be waterproof and have a more detailed app. The only user who ceased wearing it having broken it.

Fig. 6

Photo courtesy of misfit.com

Fig. 5 Averages: Set-up 4.7, Usability 5, Comfort(day) 2.5, Comfort(night) 3, Application 3.5, Rating 4, Battery 7 days, Worn 12.3 weeks

Figure 6: *Misfit Shine and Flash* (6 users) also had sync issues, but users found the technical staff at Misfit very useful. The company sent new batteries and a changing tool when the app detected a low battery. However, the lag time on this was about a month where the device was dead. The automatic sleep tracking and

waterproof quality of the device were appreciated, and it was comfortable to wear at all times. The simple data and its inflexibility was not well-liked, along with the complexity of the watch capability. A large problem came from the disk falling out of the band often, though there was the potential for the company to replace lost disks for free. Users wore it until someone else wished to try it or the device band broke or disk was lost.

Fig. 6 Averages: Set-up 4.7, Usability 5, Comfort(day) 2.5, Comfort(night) 3, Application 3.5, Rating 4, Battery 7 days, Worn 7.5 weeks

Fig. 7

Photo courtesy of getnarrative.com

Figure 7: *Narrative Clip* was not used by the (2) people who chose to take them. They found the device was too morally uncomfortable and did not want to record the people around them without their knowledge.

Figure 8: *StickNFind* (5) users found it very difficult to set up an account with the StickNFind app. They found its use to be less than expected, with a large delay in the location tracking and spontaneous beeping when close to its paired device (which was unsettling to pets.) The users were not motivated to utilize it often, though did enjoy the paging feature and found it small and unobtrusive. Users discontinued wearing them due to lack of interest or difficulty of use.

Fig. 8

Photo courtesy of sticknfind.com

Fig. 8 Averages: Set-up 4.7, Usability 5, Comfort(day) 2.5, Comfort (night) 3, Application 3.5, Rating 4, Battery 7 days, Worn 3.2 weeks

Fig. 9

Photo courtesy of jawbone.com

Figure 9: *UP24* (8 users) was easy to use, understand, and troubleshoot issues. The app was intuitive and had good data; namely recording meals and mood, setting alarms and reminders, completion percentages and sleep data. Users found it useful that times were adjustable if they had forgotten to change the mode from sleeping to awake, though they wished it could detect this automatically. The precision of its data was a concern. Users wished for a watch capability, and found it mostly comfortable, though a little bulky. They stopped wearing them due to discomfort, difficulty with use, or loss of interest.

Fig. 9 Averages: Set-up 4.7, Usability 5, Comfort(day) 2.5, Comfort(night) 3, Application 3.5, Rating 4, Battery 7 days, Worn 5.3 weeks

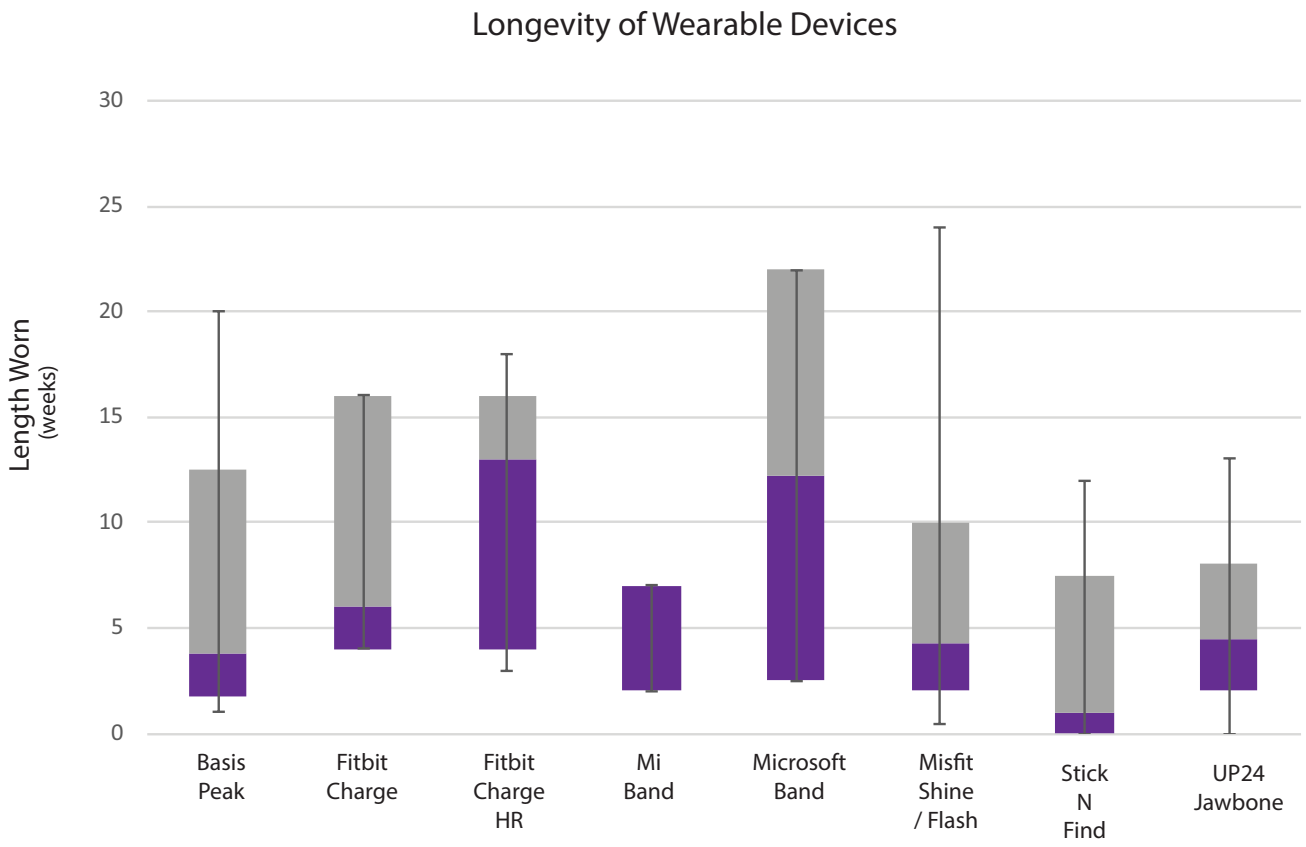


Figure 10: *Amount of Time (in weeks) Worn* Bar and Whiskers Plot based on Median. The amount of time each device was worn somewhat reflects how the user interacted with the device, how interested they were in continuing use, and how much they liked the device. A factor negating this was that people would switch devices to allow someone else to use it or to try a new device.

Discussion

From our pilot of the wearables, we have deemed the plausibility of use on a general population for the Kavli HUMAN Project study to be low, as the technology currently stands.

The obstacles that arose during the pilot that could pose a problem for the Kavli HUMAN Project included forgetfulness, loss, damage, difficulty, and change of interest. Missing data from either forgetting to put the device back on or from loss or damage of the device itself would be a large detriment on the study. Difficulty with the device, either in its physical comfort or technical use, could cause both frustration and data loss. If effort were involved in the use of the device (i.e. battery charge, remembering to put it on, fixing technical issues), interest in using the wearable is important. In the pilot, interest was relatively fleeting, with very few devices being worn for a large amount of time, and only by select people. This sustained interest was accomplished largely by the ability to see the data stream, which could not be the case in the true study.

As users were uncomfortable with using the Narrative to record during their day, it is highly unlikely a large amount of participants would utilize this sort of device in their daily life. Unstructured audio and visual

recordings may still be possible during the intake process, however. This would provide candid data of familial interactions for future analysis.

From discovering these issues, some key points were noted about giving wearables to a study population. Mainly, that effort must be minimal. This would mean a long battery life, on call technical support separate from the wearable company, and a device requiring the smallest amount of maintenance by the user. This sort of passive use would likely yield the best results. For our purposes, we would need a subtle, comfortable device that could be worn mostly undetected. This would also require the device being water-proof, to avoid forgetfulness when exiting the shower, and have a long battery life, to avoid participants needing to change the battery.

In order for the study to benefit from the wearing of the devices, participants would have to sync the devices with their phones. Ideally, automatic synching would be a feature of the device used for the Kavli HUMAN Project to maintain participant passivity. The majority of users, regardless of the type of device, left Bluetooth turned on at all times, allowing better synching. However, it was often noted that participants experienced issues synching their devices, and intervention on the part of the user would be necessary to solve these sync issues, as well as the possibility that participants who use their phones less often would need to increase the frequency of charging their phones to maintain Bluetooth use.

The subtlety of the device would be necessary due to the participants being unable to receive continuous feedback. One of the largest incentives for our pilot participants was looking at the data, which would be a detriment to the true study if allowed. As a less detrimental incentive, minimal feedback could be provided to the participants at certain intervals. The amount of feedback would have to be carefully planned as to avoid interference with the data collected. A possible attractive option would also be to allow participants to choose certain colors or styles, to match their personal preferences and increase likelihood of continued use. Another possibility would be the device having a watch capability, where the participant used it for time-telling and the data gathering aspect seemed secondary to them. This option may turn out to be highly cost exhaustive, and therefore not ideal.

The most promising device for this purpose would be the Misfit Wearables. Pilot participants found them comfortable, the devices are waterproof, and the battery life is relatively long. The largest downfall to this device being the physical structure, where the disk is highly at risk of falling out and getting lost. The Kavli HUMAN Project would require a customized device similar to this to suit our specific needs. While it is acknowledged that the feasibility of long-term use of wearables is currently low, if this ideal device were to be developed, the feasibility of longitudinal use could be more achievable.

Raw Numeric Data

Device		Ease of set-up	Usability	Comfort (day)	Comfort (sleep)	Battery Life	Application	Rating	Length Worn
Basis Peak		5	4	2	2	5	2	5	2.5
		4	4	2	3	2	2		5
		3	3				1		20
		4	5						1
		4	3						
	Average	4	3.8	2	2.5	3.5	1.7	5	7.5
Fitbit HR		1	1	3	3	5.5	3	3	16
		1	1	3	3	5	3	3	4
		1	2			5	4	4	6
	Average	1	1.3	3	3	5.2	3.8	3.3	8.7
Fitbit Charge		3	2	4	2	7	2	2	3
		1	1	5	5	5	4	4	8
		2	2	4	5	4.5	5	4	15
		1	1	1	3	7	4	1	16
		3	2	2	5	6.5	3	4	4
				3	3	12	4		13
							5		18
	Average	2	1.6	3.2	3.8	7	3.9	3	9.6
Mi Band		5	5	2	4		3		2
		5	5	3	2		4		7
		4	5						7
	Average	4.7	5	2.5	3	7	3.5	4	5.3
Microsoft Band		2	4	5	4	2	3	2	2.5
		5	3	4	4	1.5	3	4	22
		3	3	2	2				
	Average	3.3	3.3	3.7	3.3	1.75	3	3	12.3
Misfit		1	4	3	4	180	4	3	2.5
		1	1	4	4	24	3	3	24
		4	4	4	4	180	5	4	0.43
		4	5	5	5	180			2
									10
									6
	Average	2.5	3.5	4	4.3	141	4	3.3	7.5
StickNFind		3	3	5	5	1	2	3	1
		3	4	3	5	90	3	1	0
		5	2	5	3	7			12
		2	4		5				0
		1	1						3
		1							
	Average	2.5	2.8	4.3	4.5	32.7	2.5	2	3.2
UP24		5	5	4	4	7	4	4	13
		5	5	5	4	1	3	2	3
		5	4	3	3	6	4	5	8
		4	4	2	1	10	3	2	2
		5	5	2	2				2
									0
									6
									8
	Average	4.8	4.6	3.2	2.8	6	3.5	3.25	5.3

Raw Qualitative Data

From Surveys

Basis Peak

Configuring:

Easy. But the App sucks. The app is terrible and not very informative or userfriendly. Data is awesome, though.

My husband wasn't able to get this to configure on his HTC One. It was fine on my iPhone

Comfort:

This was bulky and uncomfortable to sleep with. I only used it for 4-5 days though

Missing from device:

I hate to be the scientist here, but the data presentation is generally pretty bad and they don't' do any useful statistics on it. You can't scroll through times of day easily, and it loads in a choppy way. There are actually a bug or two in the program itself. The interface isn't good. It would be great to be able to view the vector of heartrates on a much longer time scale, and to calculate means and variance through the day. It would be sweet to see trends in means on the timescale of weeks. Same with steps. There could be sweet pattern-finding algorithms as well. There's a lot of potential with these data, it's just not 100% there yet. It would be great to be able to view the vector of heartrates on a much longer time scale, and to calculate means and variance through the day. It would be sweet to see trends in means on the timescale of weeks. Same with steps. There could be sweet pattern-finding algorithms as well. There's a lot of potential with these data, it's just not 100% there yet.

I gave up on the Basis because I felt like it wasn't tracking my steps efficiently. I wore my fitbit at the same time to compare and the Basis recorded less steps at varying differences

Fitbit Charge

Configuring:

It took a while for it to sync with my phone but then it worked fine. Very user-friendly

When I first began using it, I was using the desktop interface because my phone was too old for the app. Since getting a new phone, it's much easier to check on and track my activity. Sleeping with it on for the first few nights was uncomfortable. Getting used to wearing it took some time because it's pretty bulky.

Easily paired on Android

Clearer instructions up front for future users would be very helpful, as opposed to having to research things via the Fitbit website.

Comfort:

No problems.

Needs to be a little thinner to hide under long sleeve shirt.

During the day I forget about it, and while I was fine sleeping with it at first, I find myself taking it off sometimes in the middle of the night because it's digging into something depending on how I'm sleeping.

The Fitbit is a bit bulgy for me and the bracelet is not that easy to fasten. By comparison, the jawbone was a much more discrete, seamless device. However, it is still something I'm sure some people might find ok, especially someone who is used to wearing bulky bracelets or big watches.

Missing from device:

The only way that I could keep interested is if I were getting some feedback. it's making a mess with my sleep patterns..

Nothing so far.

Activity tracking other than steps and stairs

A better food tracking system

Like on device:

FitBit website app is smooth and very user friendly, even if its measurements are coarse.

I like seeing the notifications of how many steps I achieved. It was a race between me and my husband (who wore an older model fitbit) on who would walk more.

I like the layout of the Fitbit Charge's app on the iphone. It is easy to navigate.

Weekly view

I like how it keeps all of your historical data and compares your progress

Other comments:

It thinks I'm sleeping when I'm reading. I walked 27 floors of stairs just to beat myself and I'm sweating.

I gave up on it because it was bulky and couldn't get it work properly.

Still a great device, I just still don't like sleeping with it on, which is a shame because I really like the sleep data it generates, but it's too bulky and uncomfortable for sleeping. But during the day I really enjoy knowing how much I've walked and tracking my activity.

Fitbit Charge HR

Configuring:

Configuring was plug and play, no problems other than the fact that a day or two of existing data was on the watch and impossible to delete. Usability is fine, mostly comfortable to wear and sleep with. The one notable drawback is that it isn't waterproof, meaning I have to take it off to shower or do the dishes - it would be preferable to have a simple leave-it-on-at-all-times device.

Comfort:

It is thick and snags my backpack strap when i put my backpack on. I hope I don't damage it. Otherwise I don't realize it's on as I would a watch

Bluetooth:

Only yes if I use Bluetooth. I don't use Bluetooth. This is sad cause the functionality would be much cooler if bluetooth was always on.

Missing from device:

Data analysis - easily syncing with a computer. Better data visualization.

Like on device:

Tracks heartrate overtime. The data over the long term is very good.

Other comments:

Doesn't seem to be offering much more than the app on my phone that tracks movement.

The more I use, more I get uncomfortable wearing it. Nowadays everytime I get home I just remove it from my wrist.

Still check progress and try to ensure 10,000 steps regularly, It's relatively unobtrusive, gets the job done - I'm not really looking into the data very much though. Biggest complaint is that it's not waterproof (have to take it off before a shower - easy to forget it afterwards).

Mi Band

Configuring:

Was able to set it up w/out instructions, since they were in Chinese!

Versatile, but the tracker keeps falling out of the band

Comfort:

The sleep measurements aren't very accurate, and I don't find it that comfortable for sleeping, so I take it off at night and then put it back on in the morning

Can be uncomfortably while typing on a keyboard for an extended period of time.

Like on device:

I like that the information is minimal - it flashes lights as you get closer to goal, and then vibrates when you reach goal. That's it.

Microsoft Band

Configuring:

Easy to do the initial synchronization. Hard to use the app initially... mostly because I didn't had a Microsoft account and email. In addition, the Microsoft health app is really bad.

Pretty good- but not water proof and seems much more fragile than the others (ex. easier to break, scratch, damage)

Microsoft website provides clear instruction of setting up the device step by step, but it took long time to sync and the processed was stopped many times. I didn't know it will take so long, so I was wearing the device when I synced it with smartphone and doing my work at the same time. I guess it might be the reason that the process was stopped many times since I might move my hand too far from smartphone. Plus, it takes long time to charge the buttery. Pros: The display is intuitive, easy to understand how to use the functions of the devices at first time. Customer-oriented, I'm able to change the background, the information I want to see from the device, the order of the icons, et al.

Comfort:

I've to wear it in my wrist but far away from my hand or it is uncomfortable.

It fell off my wrist last night when i was asleep…

It's think design, so absolutely uncomfortable. Especially compare to the Shine and Jawbone I have tried, it's really not a comfortable device. But after few weeks adapting with Microsoft band, I will say it's not unendurable. I'm still willing to wear it all day, if I remember to charge often.

Missing from device:

Web-application... zoom in the data, to see the activities in a better granularity

Would like a more detailed analysis for the data (ex. hourly breakdown)

Like on device:

Plot of the activity and hear rate overlapping.. This way it's easy to see both of them.

Other comments:

When I have the device on me.. I can only adjust to get it more tight.. not lighten it up.. that's a problem because I keep tightening.. It started to bother me.. then I have to remove and put it back... kind of annoying. ""I keep forgetting to set the sleep mode on. Not as easy as the fitbit. In addition, just forgot to recharge it. And for some reason my cellphone (that I recharge everyday) went out of battery at night. I think it's related to the fact that the Microsoft band was out of battery. Does the app kept trying to sync until went out of battery? Gonna test this one of those days. I forgot to remove the Band once before taking a shower, that was enough to make the visor stop working.

Misfit

Configuring:

Bluetooth is a really weak point with this - interference is a big problem - even a wireless mouse makes synching difficult. Also, in general it takes more than one try to sync this and auto synch doesn't seem to be working well.

I met a big problem with synchronizing my Shine with my iPhone 4s. I kept trying more than 1 hr, it just kept showing something like I failed and suggest me to try to unlink from another smartphone (which I did, but still didn't work). I don't think their manual is clear enough, even I checked their website, I still can't find what's the problem (I still don't know it's because Shine is insensitive or because of the problem with the Bluetooth in my iPhone 4s). Anyway, I have a no idea what I did that made synchronization work out, and so suddenly, it succeeded. I also reported this problem, and their technical staff reply my mail after 2 hr, its fast, but I have already figured out the problem. So I appreciate their service, although I still want to complain their instruction is really unclear not in manual and I don't find any informative support for troubleshooting in website. However, overall, I still likes my Shine, the synchronize processes works well after the first time. Also, the app is simple to use it (but no detail about the data and also no instruction about the meaning of their icons). To sum up, I think the biggest problem for me is I don't know what I should do when I failed to synchronize my Shine to my iPhone 4s, it makes me feel frustrated. Not knowing what I should do is worse than I don't know how to do.

First time setup was a mess but I figured out that this is likely due to the interfering registrations of multiple devices at once. The device needs Bluetooth 4.0 or higher because of the low energy mode, which also explains the problems experienced by some. Once I got home it was super easy to setup.

Comfort:

I don't generally like wearing a watch, so after a few days I switched to clipping it on.

Barely notice it

Missing from device:

More details, like my heartrate

The ability to download tracker history

Like on device:

The display is simple easy to understand. But also easy to feel bored

Actually tracks sleep without having to set start-finish time manually

Other comments:

Sync doesn't work that well - sometimes it takes a couple of tries to get the device sync'd to my phone. It's interesting and looks cool now, but I have to say at first time, I was disappointed, it just looks like cheap toy. Be honest, the real one is not so sophisticated as I expected before when I saw on website. But, it's waterproof, so I can wear it almost all day (I still put off when I was showering, but I put on immediately after that). It's soft and light, so I have no problem with wearing it while sleeping (I can't wear my normal watch all day, I usually put off my watch after I go back home immediately, so I do appreciate its light weight and comfortable material) . I got Misfit Shine, and I felt so excited first week, but now I've already feel bored. I think the reasons are: 1.Not accurate enough...it says Shine can detect my sleep automatically, but it's not correct, and the information is not useful, I still can't understand how to use that data to improve my sleep. Also, it doesn't detect my activity correctly. 2, Because its simple use App, in other words, we can't get much information about our activity, just a rough information, it makes me easy to feel bored about this device. 3. It can be used as watch, but it's inconvenient. I have to double tap my Shine, it's fun at first, but going to be inconvenient especially if my hands are occupied. These things reduce my interest in using it (although I still keep it all time) One more pros and cons about Shine are:

1. Shine's instruction is more complicated and hard to understand either by manual or website..

2. Shine's App is too simple and crude, hard to know how to use that APP in the beginning.

3. About troubleshooting, it's Hard to find the solution, especially, I remember, I had trouble with syncing and I couldn't identify the problem. But, it's fast to get respond from the company after I wrote mail to ask them how to deal with this problem.

4. Shine's also a watch, but still not convenient enough. Because I have to tap twice to know the time, sometimes, it's impossible when you wear gloves."

StickNFind

Configuring:

I like the paging feature of the app, but the location tracking is actually not as helpful as I thought it would be. I've been using it to find my keys (which are usually buried under something), and I've found that the location tracking isn't precise enough to be useful for that. There's also about ~2s delay in the movement of the dot on the screen in the app with respect to my movements, which is frustrating to use when I'm in a hurry (as I usually am, when I'm looking for my keys). I end up just paging the sticker and listening for it in order to find them.

The App was giving me issues with making an account, and it keeps saying the device hasn't been seen lately, even when it's connected

I can never log-on to the app...but once I log in its pretty fun to use

I feel the problem is the APP, it's not intuitive. First, I had problem with log in my account. I registered a new account, but it failed. I have no idea why it didn't work, no error message show up. So I registered again, it worked. Then, the second problem is I didn't know how to pair it with my smartphone, and didn't know if I succeed. No feedback message. Anyway, it's still a mystery to me, but I did succeed finally.

Comfort:

None. It's nice and tiny.

n/a - this is a Bluetooth device for a dog. It's attached to her collar, which is removed at night when she sleeps. Note that although the tracking app for the Bluetooth beacon is rarely activated, when the person who has that app is the person walking the dog, the device occasionally chirps for no apparent reason (much to the chagrin of the dog). This doesn't happen if the walker has no phone or a phone that's not paired with the device.

Be honest, I haven't tried it yet. I tied it with my keys, but I haven't lost them....so have no chance to test it. But, it's small, so I don't feel "uncomfortable" with StickNFind and not annoying at all.

Missing from device:

I still can't login to the app

More precise location tracking

Like on device:

It's alright

The page feature

Other comments:

Same old

UP24

Configuring:

The phone app is also pretty intuitive and easy to use.

The instructions are more clear, especially they provide video on website, so it's easy to follow step by step. Also, the APP is more intuitive, easy to understand the data, easy to find the icon to track data, easy to use it. About troubleshooting, it's easier to find the answer in APP or by website. (For me both experience are compare to wearing Shine)

Compared with other applications, the up24 tracks less steps.

It was fine to configure.

Comfort:

The jawbone is pretty subtle, I believe it's just a matter of getting used to wearing something on your wrist all the time. That shouldn't really be a problem with this device.

Pretty good.

Not bad, but still a little bit thick.

I'm having trouble pairing. I'm not sure that it is keeping the charge

Missing from device:

I sometimes think the data is imprecise so perhaps there should be a calibration function, where you use the device and tell it what you were doing and then based on that it gets calibrated for future data collection.

Flexibility

The only thing is it's a band so I can't know time by it, and so have to wear watch, too.

Automatic calorie calculator without entering information. I know it's not possible but it would be nice

Like on device:

I think it's good that even though I forget to set to sleep or awake mode it infers from my level of activity and then I can just go back to adjust times.

More thing I can do with Jawbone, record my daily meal, mood, set alarm, reminders…..It's more fun with using Jawbone.

Other comments:

It's nice and subtle.

About tracking, it's more correct, but can't track sleep automatically (even Shine can track sleep automatically, but Shine is less correct). Also, for me the syncing is faster than Shine.

From Email

Getting Started	Instructions
Jawbone up24	*More clear, especially they provide video on website, so it's easy to follow step by step.*
Misfit Shine	*More complicated and hard to understand either by manual or website.*

App

Jawbone up24 *More intuitive, easy to understand the data, easy to find the icon to track data, easy to use it.*

Misfit Shine *Too simple and crude, hard to know how to use that APP in the beginning*

Troubleshooting

Jawbone up24 *Easy to find the answer in APP or by website.*

Misfit Shine *Hard to find the solution, especially, I remember, I had trouble with syncing and I couldn't identify the problem. But, it's fast to get respond from the company after I wrote mail to ask them how to deal with this problem.*

Daily

Battery

Jawbone up24 *The tracker's battery last 4 days up, and it's easy and quick to charge by USB.*

Misfit Shine *Better, don't need to worry about battery*

Comfort

Jawbone up24 *Not bad, but still a little bit thick.*

Misfit Shine *Better, I can wear it all day without any problems.*

Syncing

Jawbone up24 *Faster*

Misfit Shine *Slower*

Tracking

Jawbone up24 *More correct, but can't track sleep automatically.*

Misfit Shine *Can track sleep automatically, but less correct.*

Features

Jawbone up24 *More thing I can do with Jawbone, record my daily meal, mood, set alarm, reminders…..It's more fun with using Jawbone. The only thing is it's a band so I can't know time by it, and so have to wear watch, too.*

Misfit Shine *It's also a watch, but still not convenient enough. Because I have to tap twice to know the time, sometimes, it's impossible when you wear gloves.*

Fitbit Flex

So far, I've used the following devices: Fitbit Flex for 2 months; Microsoft Band for 1 week; Basis for 2 weeks.

Between those devices, I found Flex the most comfortable, to the point that in my daily usage I hardly notice that I had it on my wrist. The Fitbit Flex is small, light, can use the device for 1 week without recharging and its waterproof. Due to these features, the device didn't affect at all my daily activities.

In comparison with Flex, Fitbit Charge makes it harder to use mouse/keyboard. With Microsoft Band and Basis, I had to remove it anytime I had to do an activity that involved little bit of water (Even wash my hands). Moreover, to get an accurate value for the heart rate measure, I had to wear the Basis and Band really tight to my wrist. It can be very uncomfortable (and doesn't work well during runs).

One cool feature that Fitbit Flex has is that it warns you when the battery is low. In my case, I configured to receive an email.

Fitbit Flex also tracks sleep. Unfortunately, I had to manually specify when I was going to sleep or when I just wake up. To start/finish this activity, I have to tap 5 times the device. Unfortunately, I would frequently forgot to start and/or finish an activity. At one point I decided to not track it anymore, since it's too much trouble to do it (tap + remember to do it) to only get a few extra information.

The Flex device presents 5 lights on it. This 'display' is used to show my current progress of the daily goal (I can manually set my daily goal). Unfortunately, this information is not very useful.

The initial synchronization between the device and the smartphone is straight forward. The Android App is clean and simple to use. Fitbit also present a website where you can see all your information. This website is easier and better to use than the App.

Pros:
- Lightweight
- Small
- Waterproof
- Comfortable
- Doesn't affect the use of keyboard/mouse

Cons:
- Start/Finish a Sleep activity is a pain
- Doesn't have any feedback from the smartphone (e.g. notification/call alert)
- Doesn't have any 'useful' display information (e.g. current time).

Comparisons:

- Charge:
> Display
> 'Automatically' track sleep

- Basis and Band:
> Track heart rate
> Display smartphone notifications

I did actually like the Fitbit charge. Key factors were having the time and caller ID linked to my phone. And the size for my 'skinny wrist' was appealing. The steps and distance seemed accurate as well. And for me it was comfortable. I think I wore it for over 5 months, nearly every day. The App and history of my activity was fine, but nothing that differentiated it from the Jawbone App I had previously used.

APPENDIX I

PUBLIC ATTITUDES ABOUT THE PMI COHORT STUDY SUMMARY

July 1, 2015

Introduction

On, July 1, 2015, David Kaufman, Program Director at National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) gave a 16 minute talk at the Precision Medicine Initiative Participant Engagement and Health Equity Workshop. The workshop took place on the NIH campus in Bethesda, Maryland, and was hosted by the Precision Medicine Initiative (PMI) Working Group of the Advisory Committee to the NIH Director (ACD). The workshop was on participant engagement and health equity as they relate to the proposed PMI national research cohort, focused on the design of an inclusive cohort, building and sustaining public trust, direct-from-participant data provision, and effective and active participant engagement characteristics of a national research cohort of one million or more volunteers. The workshop built on the big science questions developed during the April 28–29 workshop at the NIH, digital health data perspectives shared during the May 28-29 workshop, and information on the strategies to address community engagement and health disparities in a large national research cohort gathered from stakeholders through a request for information.[1]

Kaufman's talk was an overview of the results of the recent national survey on Public Attitudes about the PMI Cohort Study. The national survey was administered to determine how people feel generally about the idea of the PMI cohort, if they would be willing to participate in this sort of cohort, feelings about the cohort itself, and opinions on examples of study design for the cohort. The survey was funded by the Foundation for NIH and administered by GfK (Gesellschaft für Konsumforschung) in both Spanish and English during the span of May 28th to June 9th. The response rate for the survey was 57%, with a total of 2601 US adults participating. There was an intentional over-sampling of Black non-Hispanic (505) and Hispanic (523) people. The median time taken to complete the online survey was 21 minutes. It was noted by Kaufman that the sample for this survey was drawn from the standing GfK panel, which is known to be a fairly representative population.

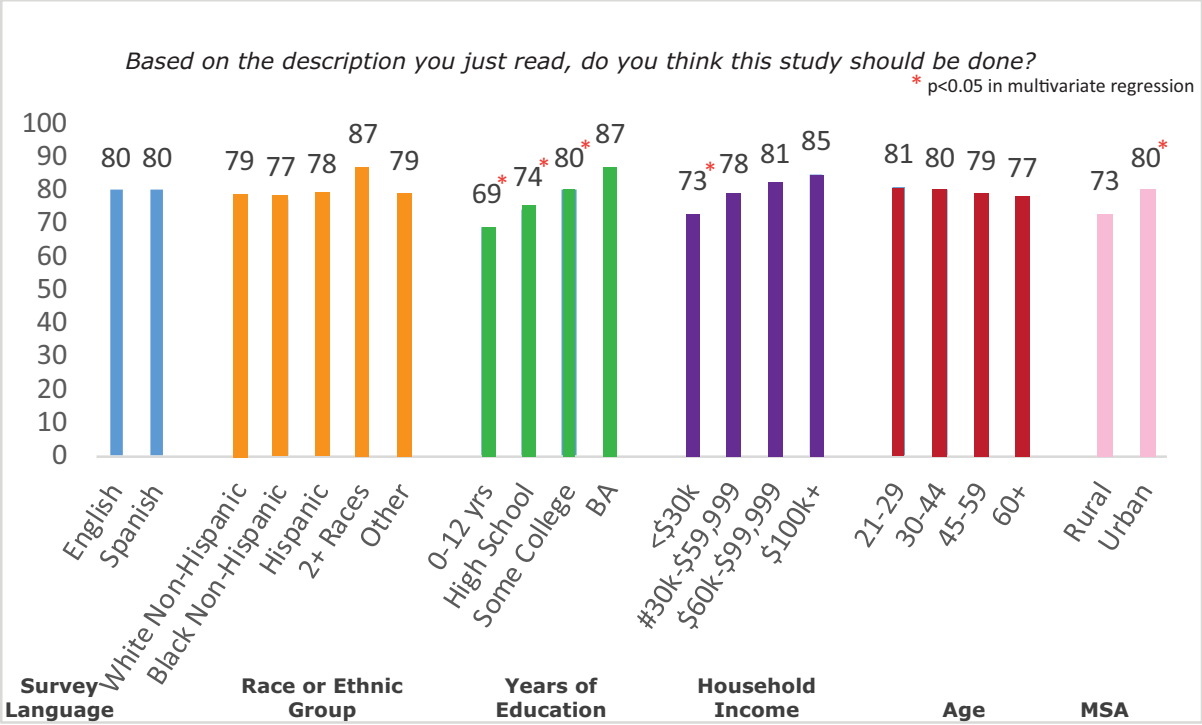
The survey's demographics were overall comparable to those of the 2010 US census, particularly in the gender distribution. The result of the over sample was seen with 7% more Black non-Hispanic and 6% more Hispanic (overall 13% less White non-Hispanic) people in the survey than in the US census. The sample was a little older than the US generally, with 10% more participants over 45 and 3% more over 65, while people in the lowest income and education brackets were underrepresented. Statistical weighting was used to correct for this, so that the results reflect what one would expect from a true representative sample of the US population.

[1] Taken from www.nih.gov. For more information and full videocasts see http://www.nih.gov/precisionmedicine/workshop-20150701.htm.

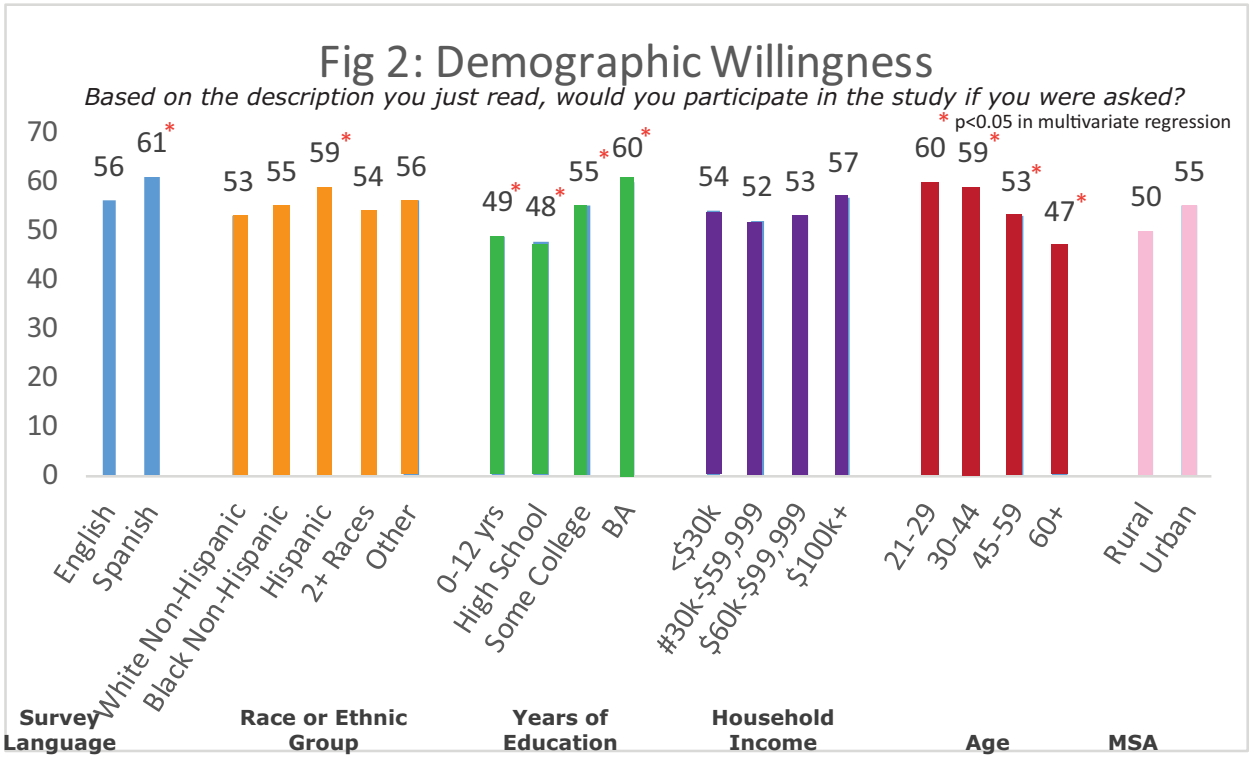
Survey Contents and Participation

At the start of the survey, participants were shown a short description of the PMI cohort study, created with as little bias as possible. They were then asked a series of questions concerning their opinions on different aspects of the study.

When asked if the study should be done, 79% of the participants said it either definitely or probably should be done, with 5% saying it definitely should not be done and 16% saying it probably should not be done. Support for the study was fairly uniform across demographics. In a regression model, support did increase with education level, people in urban areas were more supportive in general, and the lowest income level had significantly the lowest support. To determine the reason for support participants were asked about their agreement with a number of statements. Eighty percent of those who thought the study should definitely or probably be done agreed with the statement "I think the study is important to do" and 86% of those people also agreed with the statement "the study could lead to improved treatments, cures, and lives saved".



When asked if they would be willing to participate in the study if asked, 54% said they definitely or probably would, with 16% saying they definitely would not and 30% saying they probably would not. Willingness was also fairly uniform across demographics, with Hispanic people and those who took the survey in Spanish being significantly more willing to participate. The same trend with education was seen with increasing levels of education correlating to increasing willingness to participate. The opposite was seen with older participants, an increase in age showed a significant decrease in willingness to participate.

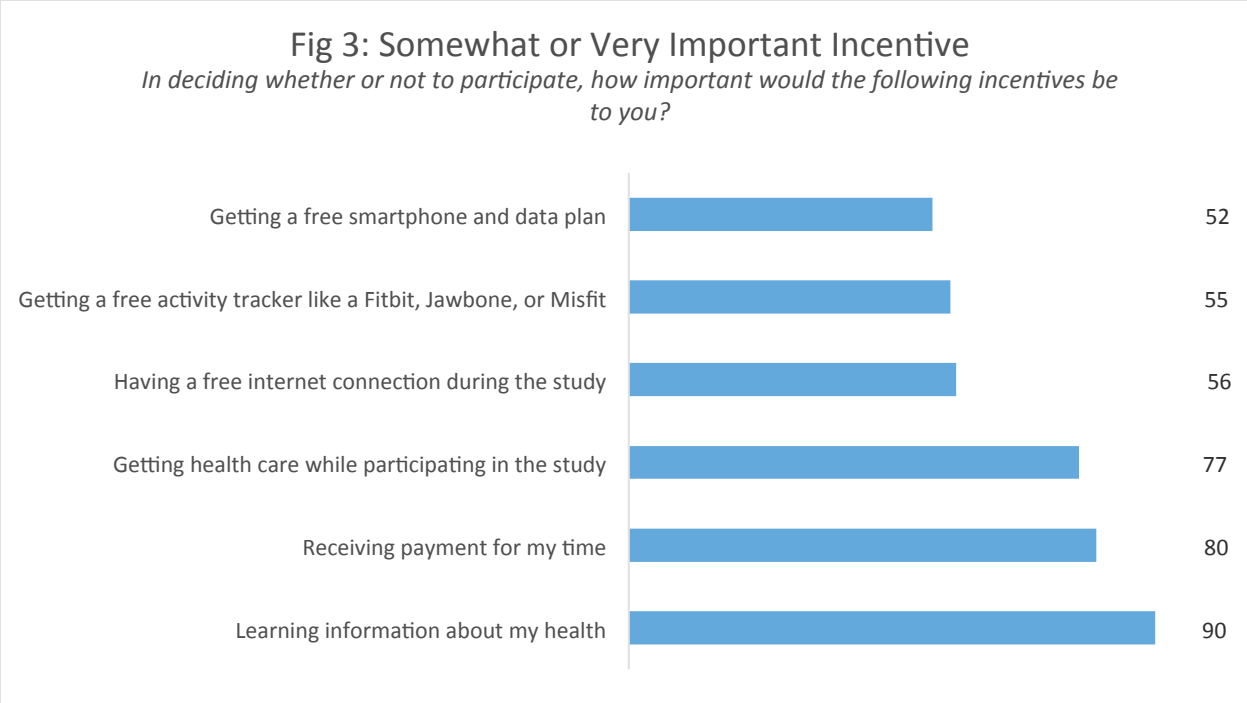


Kaufman acknowledged that even though participants say they are willing to participate in a study such as this, it does not mean that they will and that further work is necessary to make that happen. However, he remarks that these questions do suggest that no specific group seems particularly unwilling to participate.

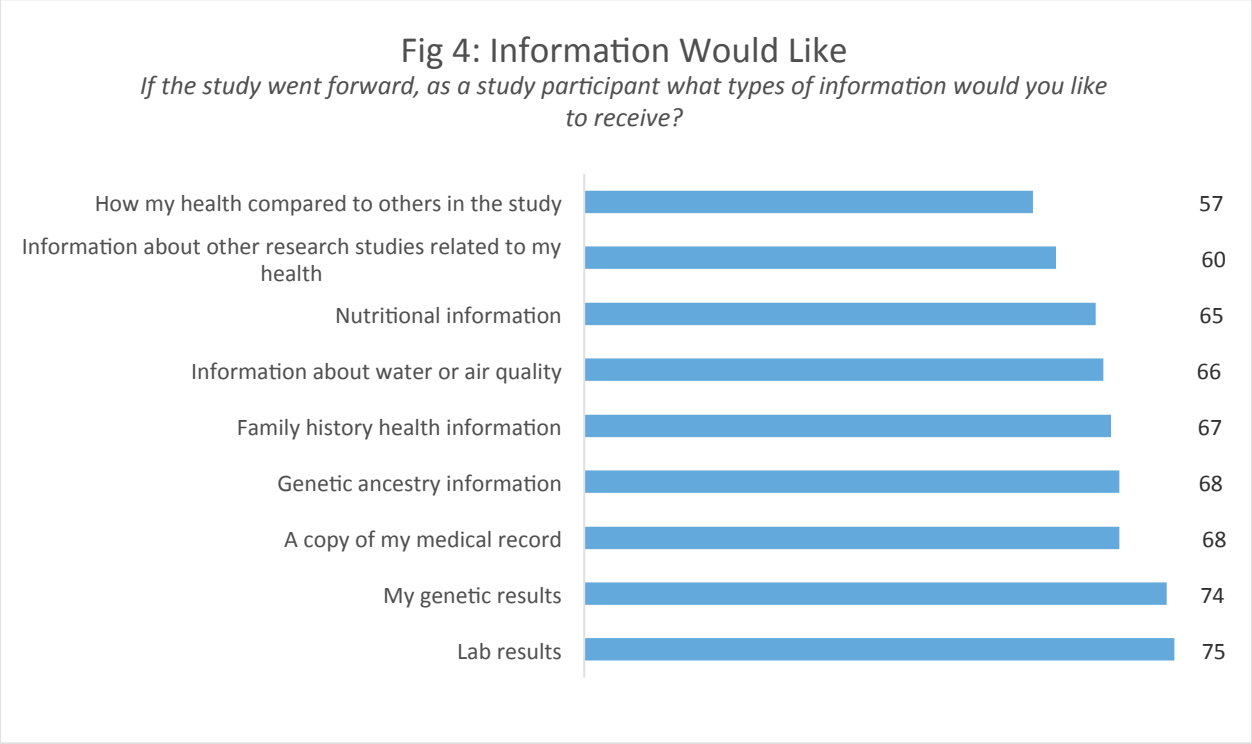
Incentives to Participate

When asked about motivations to take part in such a study, 82% agreed that "it would be interesting to receive the results of the study" while only 62% agreed that "I would take part to help advance health research." This 20% difference shows that 1 in 5 people are interested in the results, but not altruistic motivations. This means that bringing information back to participants is a very important incentive.

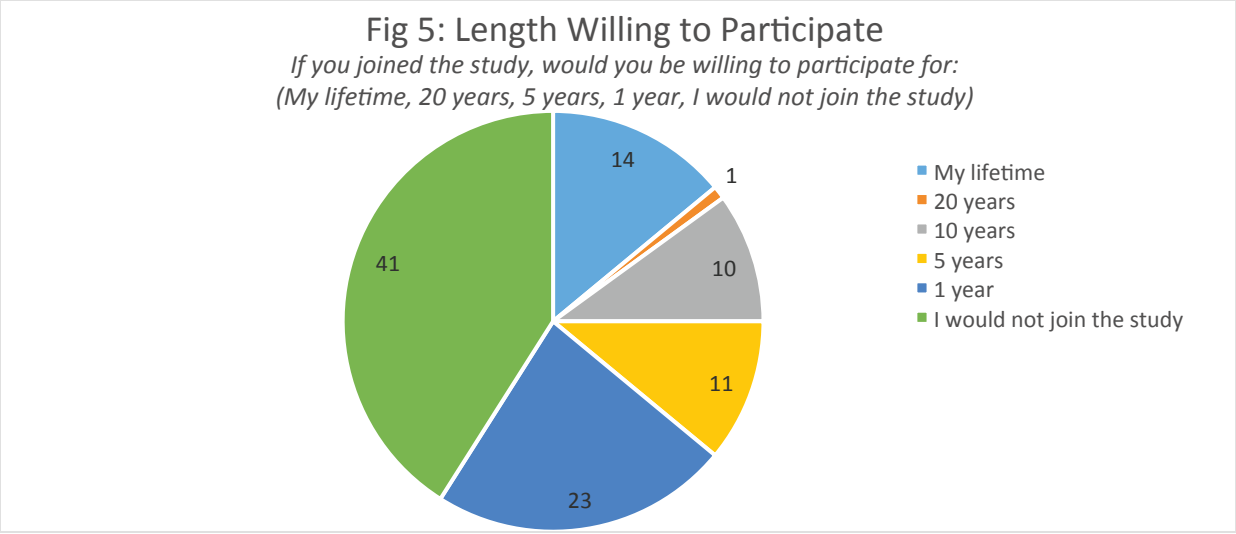
This was seen again in the ranking of given incentives, where "learning information about my health" was ranked the top incentive at somewhat or very important to 90% of the population.



When asked what kind of information they would like to receive, participants ranked lab (75%) and genetic (74%) results at the top. Another potential benefit was that this could lead to getting people into other studies, shown by 60% of responders expressed an interest in receiving "information about other research studies related to my health."

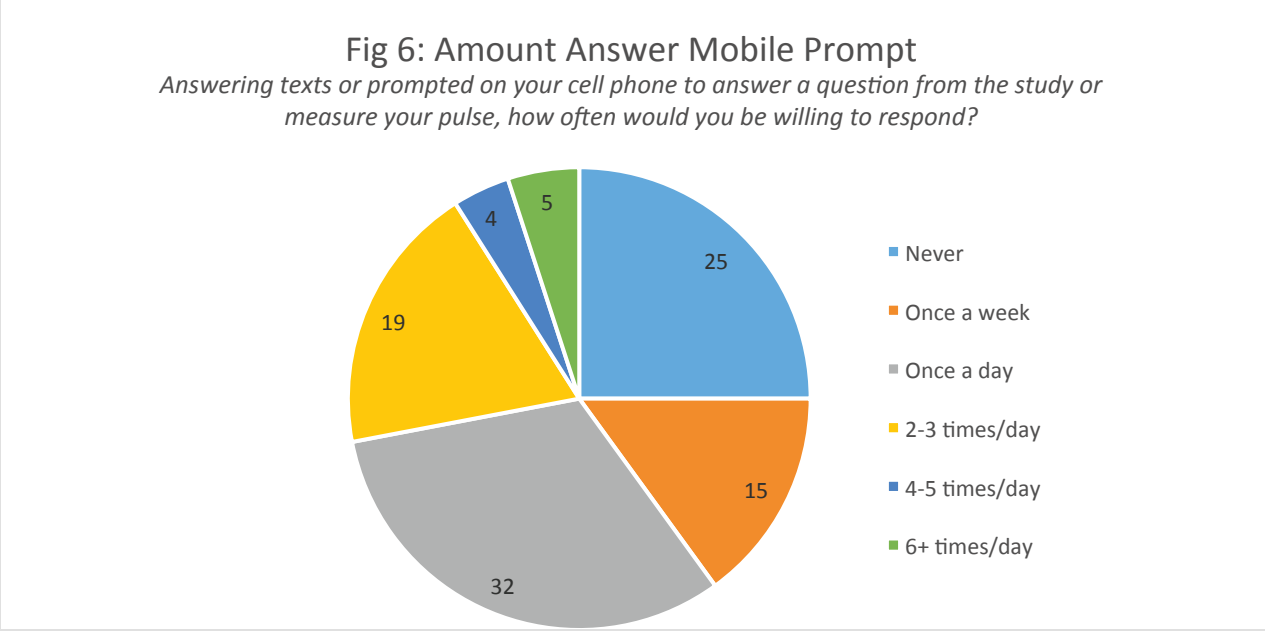


When asked how long they would remain in the study for, 41% said they were not going to join, but 25% (1 in 4) said they would stay in for 10 or more years and 36% said 5 or more years.



When asked what kinds of things they would provide, 75-85% of people said they would give lifestyle, diet, exercise, and family history information, Fitbit-like data, and soil, water, urine, hair, and saliva samples. 73% said they would give a blood sample. Only 45% said they would share social media information, about which Kaufman remarked they will have to uncover the meaning of that answer, as social media is largely publicly shared as is.

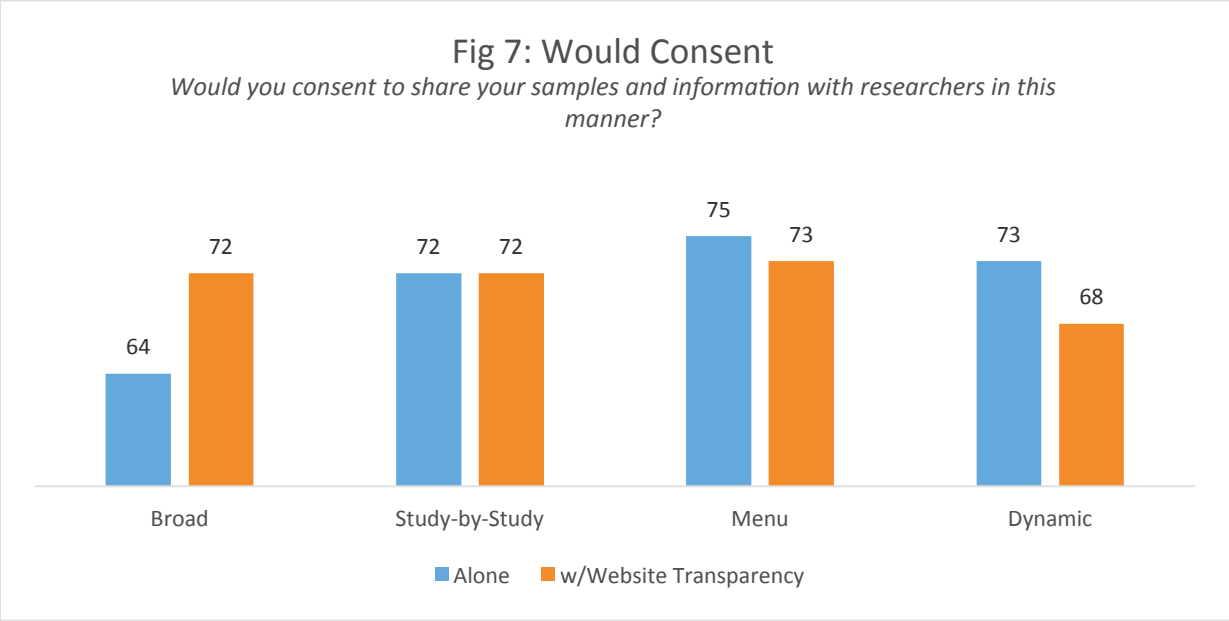
Mobile data collection was a source of interest for many, with 82% of the population having a cell or smartphone. When these people were asked how often they would partake in this sort of collection, 60% said at least once a day, while 25% said never.



Consent and Data Sharing

The survey conducted a small experiment with the sample to look at consent and data sharing options. Half of the sample was randomized to see 1 of 4 different consent models. They were shown a short description of either broad (free use of data), study-by-study (for each use), menu (you choose which categories of disease you would support at the start of the study), or dynamic (with a website to change your preferences at any time during the study) consent.

Participants showed a similar level of willingness to share information with study-by-study (72%), menu (75%), and dynamic (73%) consent. Broad consent was significantly less appealing, at 64%. The second half of the sample was also randomized and shown 1 of the 4 consent models, but with an added statement saying there would be a website where they could go to see exactly what the study was doing with their data. With this added transparency, there was no longer a difference between the consent models. Transparency is therefore very important to participants, particularly in the case of broad consent.



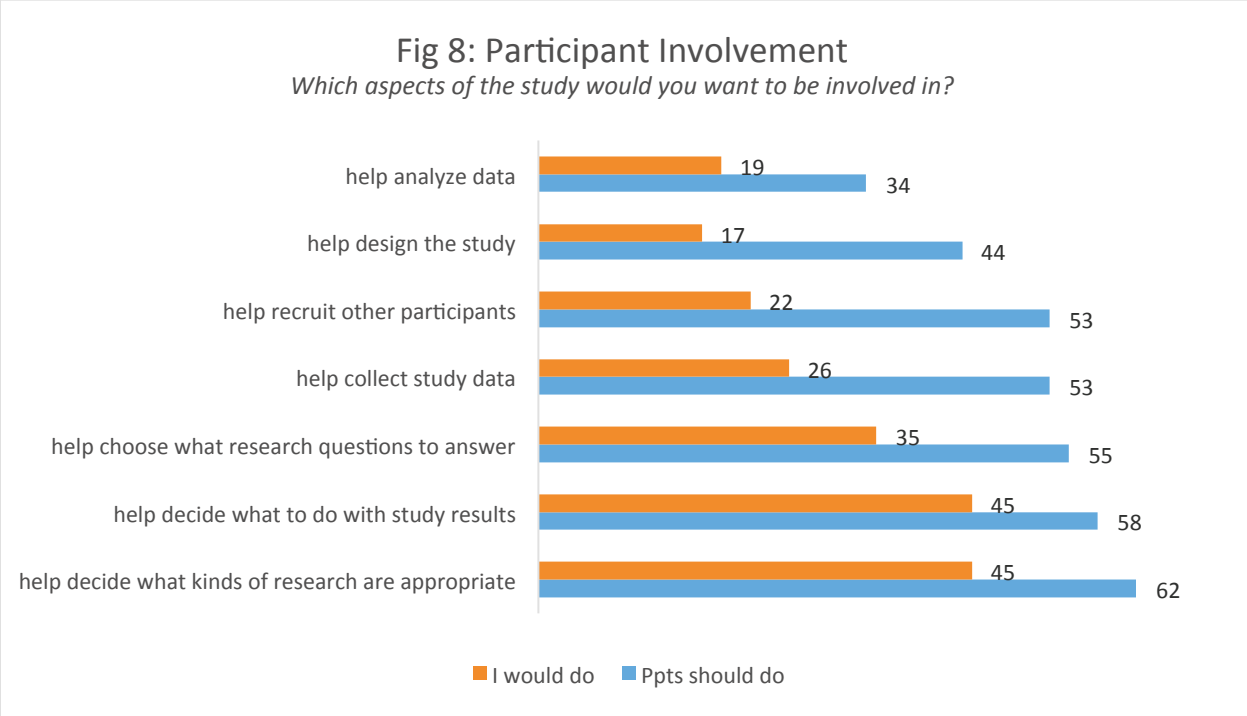
With regard to who they would be willing to let use their data, researchers at NIH had the most support at 79%, with other government researchers only having 44%. This was acknowledged to be possibly due to the vague nature of the second category, and that it may have been different with more specific organizations listed instead. Kaufman mentioned that while people may be putting more faith into the NIH researchers, they are also likely to have higher expectations and the NIH will still need to work hard to meet these. 71% would share their information with university researchers in the US, but only 39% would share it with university researchers in other countries. Kaufman noted that this being the lowest supported category has been seen before, and therefore should be looked at further. 52% would share with US pharmaceutical or drug company researchers. 43% would be willing to have their information and research results available on the Internet to anyone, if their personal information was removed first.

Participation Areas and Privacy

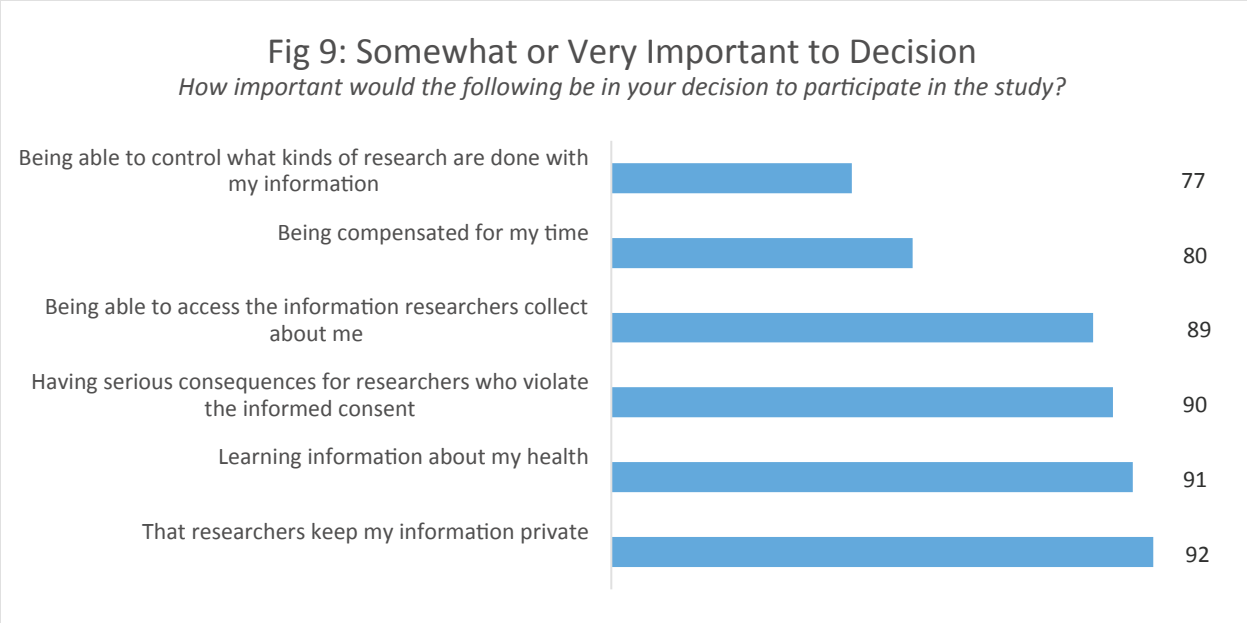
71% of participants agreed that research participants and researchers should be equal partners in the study.

In regards to what areas the sample thought participants should engage in, the top areas pertained to governance of the study, given the following options: help decide what kinds of research are appropriate, help decide what to do with the study results, help choose what research questions to answer, help collect study data, help recruit other participants, help design the study, help analyze the data. This included involvement in deciding what kind of research is appropriate, what to do with results (including shortening the feedback look and making sure results came out faster for people to be able to see them), and helping choose which research questions to answer. There was interest in other things, but less so when it came down to the details of design and analysis.

The same results were seen when the participants were asked which they personally would want to be involved in. 35-45% said they would be interested in governance aspects, and still significant numbers were interested in participating in other aspects.



When asked about design concerns, the top priority was that researchers keep information private (important to 92%). 66% of the sample agreed that the study cannot guarantee privacy, but 76% felt confident that researchers who use the data would do everything they could to keep it confidential. This shows that while participants realize there are no guarantees, they give researchers the benefit of the doubt of trying to do the right thing.



At the end of the survey, participants were asked again if they would participate in the study if asked. The responses were fairly similar to the first time the question was asked, with 56% saying definitely or probably yes, 19% saying definitely they would not, and 25% saying probably they would not. Overall, 30% of the participants shifted their category of response one way or the other, with 15% slightly more interested and 15% slightly less interested. This change shows that when people are allowed to think about the nuances and risks or benefits of a study, it helps to inform their decision.

Conclusion

Kaufman concluded that these scenarios asked about are hypothetical, and therefore the numbers are not absolute truths. However, they saw that the majority of the sample supports the idea of this PMI cohort. Willingness to participate seems relatively broad, though lower SES and older people (along with other groups) may need special attention and engagement. It will be necessary for participation overall to give health information back to participants. There is an enthusiasm for participant-partner involvement, especially in governance aspects. A broad consent model could be as acceptable as other consent models for the study, if transparency is observed and incorporated. No matter what the consent, privacy protection is critical. Related to transparency, engaging and informing possible participants will help them make informed decisions about whether to join the cohort or not.

APPENDIX J

BIOGRAPHICAL SUMMARIES

Board Members

Paul W. Glimcher

Julius Silver Professor of Neural Science, Economics and Psychology; Director, Institute for the Interdisciplinary Study of Decision Making, New York University

Paul W. Glimcher is the Julius Silver Professor of Neural Science, Economics and Psychology at New York University (NYU), Director of NYU's Institute for the Interdisciplinary Study of Decision Making and Director of the Glimcher Lab in NYU's Center for Neural Science. Dr. Glimcher pioneered the field of neuroeconomics, an interdisciplinary approach to modeling human choice and behavior. His work has made contributions to understanding how value is encoded in the brain, delay discounting and action selection in the face of both risk and ambiguity.

Andrew Caplin

Silver Professor of Economics, Department of Economics, New York University; Deputy Director, Institute for the Interdisciplinary Study of Decision Making

Andrew Caplin is a Silver Professor of Economics at New York University where he investigates new approaches to measuring and modeling individual behavior and its aggregate consequences. In addition, Dr. Caplin is a Research Associate at the National Bureau of Economic Research; Co-Principal Investigator of the Vanguard Research Initiative (a collaboration of the University of Michigan, New York University, and Vanguard); Co-Director of the Center for Experimental Social Science at NYU; and Co-Organizer of the Seminar in Neuroeconomics at NYU. He is interested in economic theory, the interface between psychology and economics, and neuroeconomics, as well as increasing returns to scale and transactions costs, household finance, and the economics of residential real estate finance. He has also testified before Congress on proposals for housing finance reform.

Lynn Goldstein

Chair, Privacy & Security Advisory Council, Kavli HUMAN Project; *Fmr. Privacy General Counsel and Chief Privacy Officer, JP Morgan Chase*

Lynn A. Goldstein is the Chair of the Kavli HUMAN Project's Privacy & Security Advisory Council and was recently the Chief Data Officer for New York University's Center for Urban Science + Progress. She joined CUSP after nearly ten years as the Chief Privacy Officer and Privacy General Counsel for JPMorgan Chase. She has also served as Chief Privacy Officer for Bank One, General Counsel for Bank One's credit card company, and Head of Litigation for Bank One, First Chicago NBD and First Chicago. Prior to these roles, Lynn was in private practice and clerked for a federal judge. Lynn is an attorney and a Certified Information Privacy Professional and a frequent speaker on privacy topics.

Gary King
Albert J. Weatherhead III University Professor, Department of Government, Harvard University

Gary King is a University Professor at Harvard University and serves as Director of the Institute for Quantitative Social Science. Dr. King develops and applies empirical methods in many areas of social science research, focusing on innovations that span the range from statistical theory to practical application. Dr. King is an Elected Fellow in 7 honorary societies, including the National Academy of Sciences, was appointed as a Senior Science Advisor to the World Health Organization and is currently a member of the Senior Editorial Board at *Science*.

Steven E. Koonin
Director, Center for Urban Science + Progress; Associate Director, Institute for the Interdisciplinary Study of Decision Making, New York University

Steven E. Koonin was appointed as the founding Director of NYU's Center for Urban Science and Progress in April 2012. That consortium of academic, corporate, and government partners will pursue research and education activities to develop and demonstrate informatics technologies for urban problems in the "living laboratory" of New York City. Prior to his NYU appointment, Dr. Koonin served as the second Under Secretary for Science at the U.S. Department of Energy where he oversaw technical activities across the Department's science, energy, and security activities and led the Department's first Quadrennial Technology Review for energy.

Julia Lane
Professor of Practice, Center for Urban Science + Progress, New York University

Julia Lane has been an Institute Fellow at American Institutes for Research, professor of economics at BETA University of Strasbourg CNRS, Chercheur, Observatoire des Sciences et des Techniques, Paris, and professor at Melbourne Institute of Applied Economics and Social Research, University of Melbourne. She was formerly director of the National Science Foundation's Science of Science and Innovation Policy program, and senior research fellow at the U.S. Census Bureau. Dr. Lane has worked with a number of national governments to document the results of their science investments, and she has testified on the topic to both the U.S. Congress and the European Parliament.

Mark Leslie
Lecturer in Management, Stanford University; Managing Director, Leslie Ventures

Mark Leslie is a successful retired entrepreneur and active member of the Silicon Valley community. He teaches courses in entrepreneurship, ethics, and sales organization. Mr. Leslie was the founding Chairman and CEO of Veritas Software, where he grew the company to 5,500 employees and a revenue base of $1.5 billion per year. He also serves on numerous boards, including Model N Corp., SugarCRMa, NYU Board of Trustees, and is chairman of the NYU Science Advisory Board.

Kathleen McGarry

Chair, Department of Economics, University of California, Los Angeles; Chair, Study Frame Advisory Council, Kavli HUMAN Project

Kathleen McGarry is a Professor of Economics and is the Chair of the Department of Economics at UCLA. She was previously the Joel Z. and Susan Hyatt, 1972 Professor of Economics at Dartmouth College and has also served on the White House Council of Economic Advisers. She has had fellowships from the Brookdale Foundation and the National Bureau of Economic Research. Dr. McGarry's research focuses on the well-being of the elderly with particular attention paid to public and private transfers, including the Medicare and Social Security Income programs, and the transfer of resources within families, especially with regard to end of life medical expenses.

Aristides A.N. Patrinos

Member, Board of Directors, Kavli HUMAN Project; Fmr. Deputy Director for Research, Center for Urban Science + Progress, New York University

Aristides Patrinos is a Member of the Kavli HUMAN Project's Board of Directors and was recently the Deputy Director for Research at New York University's Center for Urban Science + Progress. He joined CUSP from Synthetic Genomics Inc. (SGI) where he served as President and Senior Vice President for Corporate Affairs. Before SGI, Dr. Patrinos worked at the U.S. Department of Energy (DOE) in several roles, most notably, overseeing biological and environmental research in the DOE Office of Science. His accomplishments include the launch and management of the DOE's portion of the U.S. Global Change Research Program and his contributions to the Human Genome Project (HGP). Under his leadership, the DOE contributed a significant part of the first complete sequence of the human genome. Dr. Patrinos also created the DOE Joint Genome Institute and launched the Genomes to Life program.

Alex 'Sandy' Pentland

Toshiba Professor of Media Arts and Sciences; Director, Media Lab Entrepreneurship Program, Massachusetts Institute of Technology

Professor Alex 'Sandy' Pentland is a pioneer in organizational engineering, mobile information systems, and computational social science. Dr. Pentland's research focus is on harnessing information flows and incentives within social networks, the big data revolution, and converting this technology into real-world ventures. He is a lead academic adviser to the World Economic Forum, as well as founder and director of the Human Dynamics group and the Media Lab Entrepreneurship Program. Dr. Pentland is among the most-cited computer scientists in the world, and in 1997 Newsweek magazine named him one of the 100 Americans likely to shape this century.

Elizabeth A. Phelps

Julius Silver Professor of Psychology and Neural Science, Department of Psychology, New York University; Associate Director, Institute for the Interdisciplinary Study of Decision Making

Elizabeth A. Phelps studies the cognitive neuroscience of emotion, learning and memory. Her primary focus has been to understand how human learning and memory are changed by emotion and to investigate the neural systems mediating their interactions. Dr. Phelps is the recipient of the 21st Century Scientist Award from the James S. McDonnell Foundation and a fellow of the American Association for the Advancement of Science and the American Academy of Arts and Sciences.

Robert J. Shiller
Sterling Professor of Economics, Department of Economics, Yale University

Robert J. Shiller has written on financial markets, financial innovation, behavioral economics, macroeconomics, real estate, statistical methods, and on public attitudes, opinions, and moral judgments regarding markets. Dr. Shiller was awarded the Nobel Prize in Economic Sciences jointly with Eugene Fama and Lars Peter Hansen in 2013. He served as Vice President of the American Economic Association, President of the Eastern Economic Association, and was elected President of the American Economic Association for 2016. He writes regularly for Project Syndicate and The New York Times.

Miyoung Chun, ***Ex Officio***
Executive Vice President of Science Programs, The Kavli Foundation

Miyoung Chun is the Executive Vice President of Science Programs at the Kavli Foundation, where she has . Dr. Chun's academic career began as an Assistant Professor of Biochemistry and a member of Whitaker Cardiovascular Institute at Boston University School of Medicine. She subsequently worked as a scientist for Millennium Pharmaceuticals and then moved back to academia as Assistant Dean of Science and Engineering at the University of California, Santa Barbara. She has been at the Kavli foundation since 2007.

Hannah M. Bayer, ***Ex Officio***
Chief Scientist, Kavli HUMAN Project; Research Associate Professor of Decision Sciences, New York University

Hannah M. Bayer is the Chief Scientist for NYU's Institute for the Interdisciplinary Study of Decision Making (IISDM), as well as a Research Associate Professor of Decision Sciences. Dr. Bayer's work focuses on using the tools of informatics to study how people make decisions in their natural habitat, with a particular interest in the analysis of urban big data to address these research questions. She was previously a Senior Editor at Nature Neuroscience where she managed all aspects of the editorial process and journal operations.

Measurement Technology Advisory Council

Alex 'Sandy' Pentland, ***Chair***
Toshiba Professor of Media Arts and Sciences; Director, Media Lab Entrepreneurship Program, Massachusetts Institute of Technology

Professor Alex 'Sandy' Pentland is a pioneer in organizational engineering, mobile information systems, and computational social science. Dr. Pentland's research focus is on harnessing information flows and incentives within social networks, the big data revolution, and converting this technology into real-world ventures. He is a lead academic adviser to the World Economic Forum, as well as founder and director of the Human Dynamics group and the Media Lab Entrepreneurship Program. Dr. Pentland is among the most-cited computer scientists in the world, and in 1997 Newsweek magazine named him one of the 100 Americans likely to shape this century.

Nadav Aharony
Product Manager, Android, Google

Nadav Aharony is a product manager on Google's Android team, where he works on context and location-related products and technologies. He was co-founder and CEO of Behavio, a mobile sensing platform for understanding human behavior, until it was purchased by Google in 2013. Dr. Nadav has over 12 years of experience in engineering, product management, and business development roles. His work has been featured in the academic and popular press including the Wall Street Journal and Wired.

Dennis A. Ausiello
Jackson Distinguished Professor of Clinical Medicine; Director, M.D. /Ph.D. Program, Harvard Medical School & Massachusetts General Hospital

Dennis A. Ausiello the Jackson Distinguished Professor of Clinical Medicine at Harvard Medical School, and Chairman of Medicine *Emeritus* and Director of the Center for Assessment Technology and Continuous Health (CATCH) at Massachusetts General Hospital (MGH). Dr. Ausiello has made substantial contributions to the knowledge of epithelial biology in the areas of membrane protein trafficking, ion channel regulation and signal transduction. He has published numerous articles, book chapters, and textbooks and served as co-editor of *The Cecil Textbook of Medicine*. A nationally recognized leader in medicine, he is a member of the National Academies' Institute of Medicine and the American Academy of Arts and Sciences.

Jeanne Brooks-Gunn
Virginia and Leonard Marx Professor of Child Development and Education, Teachers College and College of Physicians and Surgeons, Columbia University; Co-director, National Center for Children and Families; Co-director, Columbia University Institute for Child and Family Policy

Jeanne Brooks-Gunn is a nationally-renowned scholar and expert whose research centers on family and community influences on the development of children and youth. Dr. Brooks-Gunn has also designed and evaluated interventions aimed at enhancing the well-being of children living in poverty and associated conditions. She has published over 500 articles and chapters, written 4 books, edited 13 volumes, and been the recipient of numerous major awards and honors.

Jennifer Kurkoski
Research Scientist, People Innovation Lab, Google

Jennifer Kurkoski directs Google's People Innovation Lab (PiLab) that conducts research aimed at improving the company's organizational practices. Her work has been featured in The New York Times, The Wall Street Journal, Fast Company, and Slate, as well as on the BBC and ABC's Nightline. Previously, Jennifer led community management for Excite@Home and consulted with nonprofit organizations on leadership development. Jennifer holds a Ph.D. in Business Administration (Organizational Behavior) from UC Berkeley's Haas School of Business.

David Lazer

Professor in Political Science, Computer and Information Science, Northeastern University; Visiting Scholar, Harvard University

David Lazer is Distinguished Professor of Political Science and Computer and Information Science, Northeastern University, and Co-Director, NULab for Texts, Maps, and Networks. His research focuses on the nexus of social networks, computational social science, and collaborative intelligence. He is the founder of the citizen science website "Volunteer Science." His research has been published in such journals as Science, Proceedings of the National Academy of Science, the American Political Science Review, and the Administrative Science Quarterly, and has received extensive coverage in the media, including the New York Times, NPR, the Washington Post, and CBS Evening News.

Kevin Ochsner

Professor and Director of Graduate Studies , Department of Psychology, Columbia University

Kevin Ochsner currently directs the Social Cognitive and Affective Neuroscience Laboratory at Columbia University, whose goal is to understand the inter-relationships of emotion, social behavior and self-control and their contributions to mental health and mental illness across the lifespan. Dr. Ochsner's awards include the American Psychological Association's New Investigator Award, the Cognitive Neuroscience Society's Young Investigator Award, and Columbia University's Lenfest Distinguished Faculty Award.

Roberto Rigobon

Society of Sloan Fellows Professor of Management Professor of Applied Economics, Massachusetts Institute of Technology

Roberto Rigobon is the Society of Sloan Fellows Professor of Applied Economics at the Sloan School of Management, MIT, a research associate of the National Bureau of Economic Research, a member of the Census Bureau's Scientific Advisory Committee, and a visiting professor at IESA. Dr. Rigobon focuses on the causes of balance-of-payments crises, financial crises, and the propagation of them across countries - the phenomenon that has been identified in the literature as contagion. Currently he studies properties of international pricing practices, tries to produce alternative measures of inflation, and is one of the two founding members of the Billion Prices Project, and a co-founder of PriceStats.

Benjamin Shiller

Assistant Professor of Economics, Brandeis University

Benjamin Shiller is a tenure-track assistant professor of economics at Brandeis University. His research focuses on the economic impact of digitization, with particular interest in its impact on optimal pricing, supplier coordination, and resale. Dr. Shiller's work has been featured in notable publications such as The Economist, Forbes, The Washington Post, and VOX EU. Dr. Shiller spent a year as a visiting fellow at the National Bureau of Economic Research, as part of the Economics of Digitization and Copyright Initiative.

Privacy & Security Advisory Council

Lynn Goldstein, *Chair*

Chair, Privacy & Security Advisory Council, Kavli HUMAN Project; Fmr. Privacy General Counsel and Chief Privacy Officer, JP Morgan Chase

Lynn A. Goldstein is the Chair of the Kavli HUMAN Project's Privacy & Security Advisory Council and was recently the Chief Data Officer for New York University's Center for Urban Science + Progress. She joined CUSP after nearly ten years as the Chief Privacy Officer and Privacy General Counsel for JPMorgan Chase. She has also served as Chief Privacy Officer for Bank One, General Counsel for Bank One's credit card company, and Head of Litigation for Bank One, First Chicago NBD and First Chicago. Prior to these roles, Lynn was in private practice and clerked for a federal judge. Lynn is an attorney and a Certified Information Privacy Professional and a frequent speaker on privacy topics.

Justin Brookman

Director, Consumer Privacy Project, Center for Democracy and Technology

Justin Brookman is the Director of the Center for Democracy and Technology's Consumer Privacy project. He coordinated CDT's advocacy on corporate collection, use, and retention of personal information, including efforts to enact comprehensive privacy legislation in the United States and to strengthen privacy law in Europe. Mr. Brookman has testified before House and Senate Committees on location privacy and data security, as well as the general need for stronger consumer privacy protections. He also leads CDT's work on behavioral advertising and the development of a "Do Not Track" setting for web browsers, and serves as editor of the compliance specification in the World Wide Web Consortium (W3C) standardization process.

Justin Cappos

Assistant Professor of Computer Science and Engineering, New York University

Justin Cappos is a tenure-track assistant professor in the Computer Science and Engineering department at New York University. Dr. Cappos' research philosophy focuses on improving real world systems, often by addressing issues that arise in practical deployments. His dissertation work was on Stork, the first package manager designed for environments that use operating system virtualization, such as cloud computing. Improvements in Stork, particularly relating to security, have been widely adopted and are used on the majority of Linux systems.

Marti L. Dunne

Associate Vice Provost for Research Compliance and Administration, New York University

Marti L. Dunne is the Associate Vice Provost for Research Compliance and Administration at NYU. She has a variety of offices reporting to her: the Office of Sponsored Programs,the Contracts Office, the Office of Industrial Liaison, the University Committee on Activities Involving Human Subjects (UCAIHS), the University Animal Welfare Committee (UAWC), and the Office of Veterinary Resources. In addition, her office is responsible for the development and oversight of policies concerning Conflict of Interest, Scientific Misconduct and Export Controls.

Robert M. Goerge
Senior Research Fellow, Chapin Hall, University of Chicago

Robert M. Goerge is a Senior Research Fellow at Chapin Hall at the University of Chicago. His research focuses on improving the available data and information on children and families, particularly those who require specialized services related to maltreatment, disability, poverty, or violence. Dr. Goerge developed Chapin Hall's Integrated Database on Child and Family Programs in Illinois, which links the administrative data on social service receipt, education, criminal and juvenile justice, employment, healthcare, and early childhood programs to provide a comprehensive picture of child and family use of publicly provided or financed service programs.

Thomas Hardjono
Technical Lead & Executive Director, The MIT Kerberos Consortium, Massachusetts Institute of Technology

Thomas Hardjono is the Technical Lead and Executive Director of the MIT Kerberos Consortium. He acts as a liaison with key external stakeholders who participate within MIT-KIT community, and he acts as a technical lead in establishing the strategic direction of the projects within the MIT-KIT. Dr. Hardjono has also worked on trusted computing projects in positions at Wave Systems and SignaCert. He holds 19 patents covering various security and networking technologies and has published over 50 technical papers in journals and conferences, and 3 books on security.

Jules Polonetsky
Executive Director and Co-chair, Future of Privacy Forum

Jules serves as Executive Director and Co-chair of the Future of Privacy Forum, a Washington, D.C.-based think tank that seeks to advance responsible data practices. Founded five years ago, FPF is supported by more than 80 leading companies, as well as an advisory board of comprised of the country's leading academics and advocates. FPF's current projects focus on online data use, smart grid, mobile data, big data, apps and social media.

Hilary Wandall
Associate Vice President, Compliance and Chief Privacy Officer , Merck & Co., Inc.

Hilary Wandall is Associate Vice President, Compliance and Chief Privacy Officer of Merck & Co., Inc. She has leads the Merck Privacy Office and the company's global privacy program, serves as Divisional Compliance Officer for Merck Animal Health, and is responsible for leading the global ethics and compliance program for this business unit that provides veterinary pharmaceuticals, vaccines, and health management solutions and services. She has broad multi-disciplinary experience in HIV research, genetic and cellular toxicology, internet marketing, corporate law, ethics and compliance, and privacy and data protection.

Marcy Wilder
Director, Privacy and Information Management Practice, Hogan Lovells

Marcy Wilder is a director of Hogan Lovells' global Privacy and Information Management Practice and a nationally-recognized data protection lawyer who focuses on health information law. Ms. Wilder assists clients in managing risks associated with privacy and information security practices and data breaches, including compliance with HITECH, Health Insurance Portability and Accountability Act (HIPAA), and federal and state privacy laws. Ms. Wilder served as Deputy General Counsel of the U.S. Department of Health and Human Services (HHS), where she served as the lead attorney in the development of HIPAA privacy regulations.

Miriam H. Wugmeister
Chair, Global Privacy and Data Security Group, Morrison & Foerster

Miriam H. Wugmeister is Chair of Morrison & Foerster's market-leading Global Privacy and Data Security Group. Ms. Wugmeister advises some of the world's largest and most complex multinational organizations on the planning and execution of complex global compliance efforts, assists in the negotiation of strategic deals, and defends regulatory and litigation matters relating to privacy and data security in the U.S. and internationally. Chambers USA 2014 and Chambers Global 2014 recommend Ms. Wugmeister in the top tier of U.S. privacy and data security lawyers, and Legal 500 US 2013 recognizes her as a leading lawyer for her "professionalism and strong international presence."

Study Frame Advisory Council

Kathleen McGarry, ***Chair***
Chair, Department of Economics, University of California, Los Angeles; Chair, Study Frame Advisory Council, Kavli HUMAN Project

Kathleen McGarry is a Professor of Economics and is the Chair of the Department of Economics at UCLA. She was previously the Joel Z. and Susan Hyatt, 1972 Professor of Economics at Dartmouth College and has also served on the White House Council of Economic Advisers. She has had fellowships from the Brookdale Foundation and the National Bureau of Economic Research. Dr. McGarry's research focuses on the well-being of the elderly with particular attention paid to public and private transfers, including the Medicare and Social Security Income programs, and the transfer of resources within families, especially with regard to end of life medical expenses.

BJ Casey
Director, Sackler Institute for Developmental Psychobiology; Professor of Developmental Psychobiology, Weill Medical College of Cornell University

BJ Casey is a world leader in brain imaging and its use in understanding typical and atypical development with an emphasis on understanding the brain basis of adolescent decision making. Her discoveries about the adolescent brain have significant implications for juvenile justice and mental health reform. Dr. Casey has served on several advisory boards including the National Institute for Mental Health (NIMH) Board of Scientific Counselors and NIMH Council, the Scientific Advisory Board for NARSAD, the Advisory Board for the Human Connectome Project's Lifespan Pilot Study, and National Research Council studies on assessing juvenile justice reform and sports-related concussions in youth.

Brian Elbel

Associate Professor of Population Health and Health Policy, New York University School of Medicine

Brian Elbel is an Associate Professor of Population Health and Health Policy at the NYU School of Medicine where he heads the Section on Health Choice, Policy and Evaluation within the Department of Population Health. Dr. Elbel studies how individuals make decisions that influence their health and healthcare, with a particular emphasis on evaluation, obesity and food choice. His current research includes how to use behavioral economics to influence physicians' prescribing practices; the impact of public policies mandating calorie labeling in restaurants; the impact of policies supporting the development of supermarkets in high need areas; and the influence of the food environment on childhood obesity, among others.

Arie Kapteyn

Executive Director, Dornsife Center for Economic and Social Research, University of Southern California

Arie Kapteyn is the Executive Director of the Dornsife Center for Economic and Social Research at the University of Southern California. His recent research focuses on the field of aging and economic decision making, with papers on topics related to retirement, consumption and savings, pensions and Social Security, disability, economic well-being of the elderly, and portfolio choice. He currently leads projects on several topics, including the measurement and explanation of subjective well-being, the analysis of health and economic determinants of retirement in the U.S. and Western Europe, and a center on the analysis of economic decision making related to retirement and saving and investing for retirement. He was the founding director of the CentERpanel in the Netherlands, the oldest existing probability Internet panel in the world.

Kenneth M. Langa

Professor of Medicine, University of Michigan

Kenneth M. Langa is a Professor in the Department of Internal Medicine and Institute for Social Research, a Research Scientist in the Veterans Affairs Health Services Research & Development Center for Clinical Management Research, and an Associate Director of the Institute of Gerontology, all at the University of Michigan. He is also Associate Director of the Health and Retirement Study (HRS), a National Institute on Aging-funded longitudinal study of 25,000 adults in the United States. Dr. Langa's research focuses on the epidemiology and costs of chronic disease in older adults, with an emphasis on Alzheimer's disease and other dementias. He has published more than 150 peer-reviewed articles on these topics. He is currently studying population trends in dementia prevalence, and the relationship of common cardiovascular risk factors, as well as acute illnesses, such as sepsis and stroke, to cognitive decline and dementia.

Matthew D. Lieberman

Professor of Psychology, Psychiatry and Biobehavioral Sciences; Director, Social Cognitive Neuroscience Laboratory, University of California, Los Angeles

Matthew D. Lieberman is a Professor of Psychology, Psychiatry, and Biobehavioral Sciences, and is Director of the Social Cognitive Neuroscience Laboratory, at UCLA. Dr. Lieberman, with Kevin Ochsner, coined the term "Social Cognitive Neuroscience", an area of research that integrates questions from the social sciences with methodologies of cognitive neuroscience and has become a thriving area of research. Dr. Lieberman's work has examined the neural bases of social cognition, emotion regulation, persuasion, social rejection, self-knowledge, and fairness. His research has been published in top scientific journals including *Science, American Psychologist,* and *Psychological Science*. Dr. Lieberman is also the founding editor of the journal *Social Cognitive* and *Affective Neuroscience,* and helped create the Social and Affective Neuroscience Society. In addition he won the 2007 American Psychological Association's Distinguished Scientific Award for Early Career Contribution to Psychology.

Derek Neal

Professor , Department of Economics, the Committee on Education, University of Chicago; Research Associate, National Bureau of Economic Research

Derek Neal is a Professor in the Department of Economics and the Committee on Education at the University of Chicago, and a Research Associate with the National Bureau of Economics Research. His current research focuses on the design of incentive systems for educators. His work explores the design flaws in current performance pay and accountability systems and also highlights the advantages of providing incentives through contests between schools. Dr. Neal is also exploring the causes and consequences of the prison boom in the United States. He is a recent past President of the Midwest Economics Association, a Fellow of the Society of Labor Economists, an editor of the *Journal of Political Economy*, and former Editor-in-Chief of the *Journal of Labor Economics*.

Paul Thompson

Professor of Neurology, Psychiatry, Radiology, Engineering & Ophthalmology; Director, NIH ENIGMA "Big Data" Center of Excellence ; Associate Dean for Research, Keck USC School of Medicine
Director, USC Imaging Genetics Center
University of Southern California

Paul Thompson is a Professor of Neurology, Psychiatry, Radiology, Engineering & Ophthalmology at the University of Southern California (USC). Dr. Thompson also directs the NIH ENIGMA Consortium, an NIH big data center of excellent and global alliance of 307 scientists in 33 countries who conduct the largest studies of 10 major brain diseases—ranging from schizophrenia, depression, ADHD, bipolar illness and OCD, to HIV and addictions of the brain. Among his many accomplishments, Dr. Thompson's team created the first maps of Alzheimer's disease and schizophrenia spreading in the living brain, and a method to track brain growth in children.

Scientific Agenda Advisory Council

Andrew Caplin, ***Chair***
Silver Professor of Economics, Department of Economics, New York University; Deputy Director, Institute for the Interdisciplinary Study of Decision Making

Andrew Caplin is a Silver Professor of Economics at New York University where he investigates new approaches to measuring and modeling individual behavior and its aggregate consequences. In addition, Dr. Caplin is a Research Associate at the National Bureau of Economic Research; Co-Principal Investigator of the Vanguard Research Initiative (a collaboration of the University of Michigan, New York University, and Vanguard); Co-Director of the Center for Experimental Social Science at NYU; and Co-Organizer of the Seminar in Neuroeconomics at NYU. He is interested in economic theory, the interface between psychology and economics, and neuroeconomics, as well as increasing returns to scale and transactions costs, household finance, and the economics of residential real estate finance. He has also testified before Congress on proposals for housing finance reform.

Dennis A. Ausiello
Jackson Distinguished Professor of Clinical Medicine
Director, M.D. /Ph.D. Program, Harvard Medical School & Massachusetts General Hospital

Dennis A. Ausiello the Jackson Distinguished Professor of Clinical Medicine at Harvard Medical School, and Chairman of Medicine *Emeritus* and Director of the Center for Assessment Technology and Continuous Health (CATCH) at Massachusetts General Hospital (MGH). Dr. Ausiello has made substantial contributions to the knowledge of epithelial biology in the areas of membrane protein trafficking, ion channel regulation and signal transduction. He has published numerous articles, book chapters, and textbooks and served as co-editor of *The Cecil Textbook of Medicine*. A nationally recognized leader in medicine, he is a member of the National Academies' Institute of Medicine and the American Academy of Arts and Sciences.

Laura Jean Bierut
Alumni Endowed Professor of Psychiatry; Co-Director, Outpatient Psychiatry Clinic, Washington University School of Medicine

Laura Jean Bierut the Alumni Endowed Professor of Psychiatry and Co-Director of the Outpatient Psychiatric Clinic at Washington University School of Medicine. Dr. Bierut is a physician-scientist with significant experience in genetic studies of smoking behaviors, addiction, and other psychiatric and medical illnesses. She is one of 14 investigators who received funding through the National Human Genome Research Institute's Genes, Environment and Health Initiative, and she led the effort to understand the interplay of genes and environment in the development of addiction. She is an active member in the National Institute on Drug Abuse's (NIDA) Genetics Consortium, a national group of scientists who are leading NIDA's efforts to understand genetic causes of substance dependence. She also leads The Collaborative Genetic Study of Nicotine Dependence (COGEND).

Clancy Blair

Professor of Cognitive Psychology, Department of Applied Psychology, , Steinhardt School of Culture, Education, and Human Development, New York University

Clancy Blair is a Professor of Cognitive Psychology at New York University where he focuses on developmental psychology and studies self-regulation in young children. His primary interest concerns the development of cognitive abilities referred to as "executive functions" and the ways in which these aspects of cognition are important for school readiness and early scholastic achievement. He is also interested in the development and evaluation of preschool and elementary school curricula designed to promote executive functions as a means of preventing academic failure. In 2002, Dr. Blair and his colleagues at Pennsylvania State University (Penn State) and the University of North Carolina at Chapel Hill received funding from the National Institute of Child Health and Human Development for a longitudinal, population-based study of family ecology and child development beginning at birth. For the project Dr. Blair is examining the impact of early experiential and biological influences on the development of executive functions, the interactions between those influences, and related aspects of self-regulation.

Jeanne Brooks-Gunn

Virginia and Leonard Marx Professor of Child Development and Education, Teachers College and College of Physicians and Surgeons, Columbia University; Co-director, National Center for Children and Families; Co-director, Columbia University Institute for Child and Family Policy

Jeanne Brooks-Gunn is a nationally-renowned scholar and expert whose research centers on family and community influences on the development of children and youth. Dr. Brooks-Gunn has also designed and evaluated interventions aimed at enhancing the well-being of children living in poverty and associated conditions. She has published over 500 articles and chapters, written 4 books, edited 13 volumes, and been the recipient of numerous major awards and honors.

BJ Casey

Director, Sackler Institute for Developmental Psychobiology; Professor of Developmental Psychobiology, Weill Medical College of Cornell University

BJ Casey is a world leader in brain imaging and its use in understanding typical and atypical development with an emphasis on understanding the brain basis of adolescent decision making. Her discoveries about the adolescent brain have significant implications for juvenile justice and mental health reform. Dr. Casey has served on several advisory boards including the National Institute for Mental Health (NIMH) Board of Scientific Counselors and NIMH Council, the Scientific Advisory Board for NARSAD, the Advisory Board for the Human Connectome Project's Lifespan Pilot Study, and National Research Council studies on assessing juvenile justice reform and sports-related concussions in youth.

David Cesarini

Assistant Professor of Economics, Department of Economics, Center for Experimental Social Science, New York University

David Cesarini is an Assistant Professor in the Department of Economics & Center for Experimental Social Science at New York University. He is an empirically oriented economist with interests in applied microeconomics, as well as behavioral and experimental economics. Much of Dr. Cesarini's work has

used genetically informative datasets, often coupled with experimental methods, to answer questions about sources of individual differences in economic preferences, behaviors, and outcomes. In recent work he has explored how molecular genetic data can be used to shed light on economic questions.

David M. Cutler

Harvard College Professor, Otto Eckstein Professor of Applied Economics, Harvard University

David Cutler is a Harvard College Professor and Otto Ecksteiin Professor of Applied Economics at Harvard University with secondary appointments at the Kennedy School of Government and School of Public Health. He is also a Research Associate at the National Bureau of Economic Research, a member of the National Academies' Institute of Medicine, and a Fellow of the Employee Benefit Research Institute. Professor Cutler served on the Council of Economic Advisers and the National Economic Council during the Clinton Administration and has advised the presidential campaigns of Bill Bradley and John Kerry, as well as being Senior Health Care Advisor for the Obama Presidential Campaign.

Adam Drewnowski

Professor, Epidemiology; Director, Nutritional Sciences Program, University of Washington

Adam Drewnowski is a Professor of epidemiology and Director of the Nutritional Sciences Program at the University of Washington. He is a world-renowned leader in innovative research approaches for the prevention and treatment of obesity. Dr. Drewnowski is Director of the Center for Public Health Nutrition and the Center for Obesity Research. Dr. Drewnowski's current research focuses on the relationship between poverty and obesity and the links between obesity and diabetes rates in vulnerable populations and access to healthy foods. He has also conducted epidemiological studies on dietary quality, both in the U.S. and abroad, and extensive research on taste function and food preferences in relation to food choices and the overall quality of the diet.

Edward Glaeser

Fred and Eleanor Glimp Professor of Economics, Harvard University

Edward Glaeser is the Fred and Eleanor Glimp Professor of Economics at Harvard University, a Senior Fellow at the Manhattan Institute, contributing editor of *City Journal*, and a contributor to *The New York Times*' *Economix* blog. Professor Glaeser teaches urban and social economics and microeconomic theory. He has published dozens of papers on cities, economic growth, and law and economics. In particular, his work has focused on the determinants of city growth and the role of cities as centers of idea transmission. Glaeser also edits the *Quarterly Journal of Economics*. His book, *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier* (Penguin Press) was published in 2011.

Masoud Ghandehari

Head, Urban Observatory, Center for Urban Science + Progress; Associate Professor, Civil & Environmental Engineering, Polytechnic School of Engineering, New York University

Masoud Ghandehari is an Associate Professor of Civil & Environmental Engineering in New York University's Polytechnic School of Engineering and is the Head of the Urban Observatory at NYU's Center for Urban Science + Progress. Dr. Ghandehari's academic research has focused on the understanding of the performance, aging and health of civil infrastructure, and the application of optical

methods for materials diagnostics and environmental sensing. Through the application of sensing, observations, and system assessment, he is developing methodologies that generate multi-scale urban data on the physical, environmental and human systems in cities. This work is aimed at developing novel approaches to understanding the condition and well-being of cities and inhabitants. Dr. Ghandehari is the founding investigator of the New York State Resiliency Institute for Storm Events (NYRISE), and founder of Chromosense LLC—supported by the National Institute of Health—for innovation in environmental sensing.

Pamela Giustinelli

Research Assistant Professor, University of Michigan Institute for Social Research

Pamela Giustinelli is a Research Assistant Professor at the University of Michigan's Institute for Social Research and affiliated with the Michigan Center on Demography of Aging, the Center for European Studies, and the Michigan Institute for Data Science. He is also a member of the Human Capital and Economic Opportunity Global Working Group of the University of Chicago. She is interested in modeling, empirical, and counterfactual policy analysis of individual and multilateral decision making under uncertainty-ambiguity, especially as it applies to the family and human capital contexts, and how survey methodology particularly relates to this line of research. Her work examines the relative roles played by preferences, beliefs, choice sets, and other elements of decision-making processes in both the selection of decision rules and outcomes.

Catherine Hartley

Assistant Professor of Psychology in Psychiatry Sackler Institute for Developmental Psychobiology, Weill Medical College of Cornell University

Catherine Hartley is an Assistant Professor of Psychology in Psychiatry at Weill Medical College of Cornell University. Her research examines the diverse forms of learning used to assign value to stimuli and actions, and how these learning processes influence decision-making and psychological well-being. Dr. Hartley uses a range of techniques, including structural and functional neuroimaging, psychophysiology, computational modeling, genetics, and behavioral learning and decision making paradigms.

Ichiro Kawachi

John L. Loeb and Frances Lehman Loeb Professor of Social Epidemiology; Chair, Department of Social and Behavioral Sciences, Harvard University

Ichiro Kawachi is the John L. Loeb and Frances Lehman Loeb Professor of Social Epidemiology, and Chair of the Department of Social and Behavioral Sciences at Harvard University. He is also the Co-Director of the Robert Wood Johnson Foundation Health and Society Scholars, co-Director of the Initiative to Maximize Student Diversity Training Grant, chair of the Harvard School of Public Health's Institutional Review Board, and co-Editor-in-Chief of the international journal *Social Science & Medicine*. His current NIH-funded project is focused on the longitudinal impacts of community social cohesion/social capital on functional recovery after the March 11, 2011 Great Eastern Japan earthquake and tsunami. In 2013, he launched a massive, open online course (MOOC) through HarvardX called "Health and Society" (PHx 201), in which 32,000 participants registered worldwide.

Kenneth M. Langa
Professor of Medicine , University of Michigan

Kenneth M. Langa is a Professor in the Department of Internal Medicine and Institute for Social Research, a Research Scientist in the Veterans Affairs Health Services Research & Development Center for Clinical Management Research, and an Associate Director of the Institute of Gerontology, all at the University of Michigan. He is also Associate Director of the Health and Retirement Study (HRS), a National Institute on Aging-funded longitudinal study of 25,000 adults in the United States. Dr. Langa's research focuses on the epidemiology and costs of chronic disease in older adults, with an emphasis on Alzheimer's disease and other dementias. He has published more than 150 peer-reviewed articles on these topics. He is currently studying population trends in dementia prevalence, and the relationship of common cardiovascular risk factors, as well as acute illnesses, such as sepsis and stroke, to cognitive decline and dementia.

Scott Lipnick
Scientific Director, Center for Assessment Technology and Continuous Health (CATCH); Assistant in Biomedical Physics, Department of Medicine, Massachusetts General Hospital
Imaging and Data Specialist, Stem Cell and Regenerative Biology Department, Harvard University

Scott Lipnick is the Scientific Director of the Center for Assessment Technology and Continuous Health (CATCH) at Massachusetts General Hospital, as well as an Assistant in Biomedical Physics in the Department of Medicine at Mass General. He also holds an associate position at Harvard University in the Stem Cell and Regenerative Biology Department as an Imaging and Data Specialist. Dr. Lipnick most recently served as the Director of Scientific Programs at the New York Stem Cell Foundation Research Institute, where his efforts focused on developing new tools for scaling up operations. Prior to this, he served as a Science & Technology Policy Fellow of the American Association for the Advancement of Science (AAAS) at the National Institutes of Health's Center for Regenerative Medicine (NIH CRM).

Charles F. Manski
Board of Trustees Professor in Economics, Northwestern University

Charles Manski research focuses on econometrics, judgement and decision, and the analysis of social policy. Previous appointments include Hilldale Professor at the University of Wisconsin-Madison and Associate Professor at the Hebrew University of Jerusalem. He has also served as Director of the Institute for Research on Poverty and as Chair of the Board of Overseers of the Panel Study of Income Dynamics (PSID). He is also a Member of the National Academy of Sciences, a Fellow of the Econometric Society, the American Academy of Arts and Sciences, and the American Association for the Advancement of Science, and a Corresponding Fellow of the British Academy. Past and current service at the National Research Council includes being Chair of the Committee on Data and Research for Policy on Illegal Drugs and a member of the Report Review Committee and the Commission on Behavioral and Social Sciences and Education.

Kathleen McGarry

Chair, Department of Economics, University of California, Los Angeles; Chair, Study Frame Advisory Council, Kavli HUMAN Project

Kathleen McGarry is a Professor of Economics and is the Chair of the Department of Economics at UCLA. She was previously the Joel Z. and Susan Hyatt, 1972 Professor of Economics at Dartmouth College and has also served on the White House Council of Economic Advisers. She has had fellowships from the Brookdale Foundation and the National Bureau of Economic Research. Dr. McGarry's research focuses on the well-being of the elderly with particular attention paid to public and private transfers, including the Medicare and Social Security Income programs, and the transfer of resources within families, especially with regard to end of life medical expenses.

Aristides A.N. Patrinos

Member, Board of Directors, Kavli HUMAN Project; Fmr. Deputy Director for Research, Center for Urban Science + Progress, New York University

Aristides Patrinos is a Member of the Kavli HUMAN Project's Board of Directors and was recently the Deputy Director for Research at New York University's Center for Urban Science + Progress. He joined CUSP from Synthetic Genomics Inc. (SGI) where he served as President and Senior Vice President for Corporate Affairs. Before SGI, Dr. Patrinos worked at the U.S. Department of Energy (DOE) in several roles, most notably, overseeing biological and environmental research in the DOE Office of Science. His accomplishments include the launch and management of the DOE's portion of the U.S. Global Change Research Program and his contributions to the Human Genome Project (HGP). Under his leadership, the DOE contributed a significant part of the first complete sequence of the human genome. Dr. Patrinos also created the DOE Joint Genome Institute and launched the Genomes to Life program.

Russell Poldrack

Professor, Psychology; Member, Stanford Neurosciences Institute, Stanford University

Russell Poldrack is a Professor of Psychology at Stanford University where he researches decision science, learning and memory, motivation and emotion, plasticity and change, and psychopathology and risk, while employing computational, developmental, and neuroscience approaches. Dr. Poldrack uses a range of neuroimaging and behavioral techniques to investigate the organization of cognitive and neural systems involved in learning and memory, decision making, and executive control. His laboratory also develops novel informatics and data analysis approaches and tools to improve extraction and synthesis of knowledge about the structure of mind-brain relations. Dr. Poldrack is the founding Editor-in-Chief of *Frontiers in Brain Imaging Methods* and ad hoc Handling Editor for the *Proceedings of the National Academy of Sciences*, among others.

C. Cybele Raver

Vice Provost for Academic, Research, and Faculty Affairs, New York University

C. Cybele Raver is the Vice Provost for Research and Faculty Affairs at New York University. Prior to joining the Provost's Office, Dr. Raver served as inaugural director of NYU's Institute of Human Development and Social Change (IHDSC). As a behavioral social scientist trained in psychology and public policy, Dr. Raver played a key role in fostering interdisciplinary research at NYU through the IHDSC. Dr. Raver's own program of research focuses on early learning and development in the contexts

of poverty and policy. She also examines the mechanisms that support children's cognitive and emotional outcomes in the context of early educational intervention. Dr. Raver and her research team currently oversee the Chicago School Readiness Project, a federally-funded longitudinal study of the short- and long-term impacts of preschool intervention for low-income children in Chicago. Dr. Raver also serves as a co-investigator on several other large educational evaluation studies. In addition to her work at NYU, Dr. Raver regularly advises local and federal government agencies and foundations on promoting healthy development and learning among children from birth to the 3rd grade. Her research has garnered several prestigious awards from organizations such as the American Psychological Association and the William T. Grant Foundation. Dr.

Regina Sullivan

Professor, Department of Child and Adolescent Psychiatry, Child and Adolescent Psychiatry, New York University

Regina Sullivan is a Professor of Child and Adolescent psychiatry and a Developmental Behavioral Neuroscientist with NYU's Langone School of Medicine. Her research has highlighted how the infant brain functions differently from the adult brain, as well as the critical role of the caregiver in modifying how the young brain responds to trauma and elucidating the neural mechanisms for the enduring mental health effects of abuse and trauma in early life. She also serves on boards for scientific journals including, *Developmental Cognitive Neuroscience Journal*, *International Journal for Developmental Psychobiology*, and *Frontiers in Behavioral Neuroscience*.

Cass R. Sunstein

Robert Walmsley University Professor, Harvard Law School

Cass Sunstein is the Robert Walmsley University Professor at Harvard Law School. Dr. Sunstein is currently working on group decision making and various projects on the idea of liberty. Previously, Dr. Sunstein was the Administrator of the White House Office of Information and Regulatory Affairs from 2009 to 2012. Mr. Sunstein has testified before congressional committees on many subjects and has been involved in constitution-forming and law reform activities in a number of nations. He is also a Bloomberg View columnist and a member of the Bloomberg Government Advisory Board. He writes widely on topics ranging from behavioral economics to constitutional, administrative, and environmental law. His interests include constitutional law, administrative law, environmental law, and law and behavioral economics.

Scott Schuh

Director, Consumer Payments Research Center; Senior Economist and Policy Advisor, Research Department, Federal Reserve Bank of Boston

Scott Schuh is Director of the Consumer Payments Research Center (CPRC) and a Senior Economist and Policy Advisor in the Research Department of the Federal Reserve Bank of Boston. He joined the Bank in 1997 after serving as an economist at the Board of Governors of the Federal Reserve System since 1991. Dr. Schuh also worked for President Reagan's Council of Economic Advisers, the U.S. Congressional Budget Office, and the U.S. Census Bureau. Dr. Schuh's current research focuses on consumer choices pertaining to money, payments, and banking, and on leading the CPRC to produce the annual *Survey of Consumer Payment Choice* and a new *Diary of Consumer Payment Choice*. His earlier research is in macroeconomics, labor economics, and international economics. A distinctive feature of Dr. Schuh's research agenda has been a focus on the microeconomic foundations of macroeconomic fluctuations and

growth. He co-authored two books, including the award-winning *Job Creation and Job Destruction* (1995), and has published articles in journals such as the *Journal of Monetary Economics; International Economic Review; Journal of Money, Credit, and Banking; Review of Economic Dynamics;* and *Journal of International Economics*.

Wolfram Schultz

Wellcome Principal Research Fellow, Professor of Neuroscience, University of Cambridge

Wolfram Schultz is the Wellcome Principal Research Fellow and Professor of Neuroscience at the University of Cambridge. His research interests include behavioral analysis, electrophysiological recording techniques, and functional magnetic resonance imaging (fMRI). Dr. Schultz is working to relate the mechanics of brain activity to measurable behavior. He combines neurophysiological, imaging, and behavioral techniques to investigate the neural correlates of goal-directed behavior. During his career, Dr. Schultz has won a number of awards, including the 1984 Ellermann Price of the Swiss Societies for Neurology, Neurosurgery and Neuropathology and the 2005 Ipsen Prize for Neuronal Plasticity. He also serves on the Editorial Board of the *Journal of Neurophysiology* and is an Associate Editor of the *Proceedings of the Royal Society*.

Robert M. Townsend

Elizabeth and James Killian Professor of Economics
Department of Economics, Massachusetts Institute of Technology

Robert M. Townsend is the Elizabeth and James Killian Professor of Economics at MIT. Prior to joining MIT, he was the Charles E. Merriam Distinguished Service Professor in the Department of Economics at the University of Chicago, where he remains a Research Professor. Dr. Townsend is a theorist, macroeconomist, and development economist who analyzes the role and impact of economic organization and financial systems through applied general equilibrium models, contract theory and the use of micro data. He is known for his seminal work on costly state verification, the revelation principle, optimal multi-period contracts, decentralization of economies with private information, models of money with spatially separated agents, forecasting the forecasts of others, and insurance and credit in developing countries.

APPENDIX K

WHITE PAPERS

As described in Scientific Agenda section of the design document, 22 white papers have been commissioned as of October 2015. Five have been published, one is in review, 16 are in progress, and one was recently invited. This appendix contains the first published set followed by drafts of the six in progress papers.

Big Data
Volume 3 Number 3, 2015
DOI: 10.1089/big.2015.0012

ORIGINAL ARTICLE

Using Big Data to Understand the Human Condition: The Kavli Human Project

Okan Azmak,[1] Hannah Bayer,[1] Andrew Caplin,[1,*] Miyoung Chun,[2] Paul Glimcher,[1] Steven Koonin,[1] and Aristides Patrinos[1]

Abstract

Until now, most large-scale studies of humans have either focused on very specific domains of inquiry or have relied on between-subjects approaches. While these previous studies have been invaluable for revealing important biological factors in cardiac health or social factors in retirement choices, no single repository contains anything like a complete record of the health, education, genetics, environmental, and lifestyle profiles of a large group of individuals at the within-subject level. This seems critical today because emerging evidence about the dynamic interplay between biology, behavior, and the environment point to a pressing need for just the kind of large-scale, long-term synoptic dataset that does not yet exist at the within-subject level. At the same time that the need for such a dataset is becoming clear, there is also growing evidence that just such a synoptic dataset may now be obtainable—at least at moderate scale—using contemporary big data approaches. To this end, we introduce the Kavli HUMAN Project (KHP), an effort to aggregate data from 2,500 New York City households in all five boroughs (roughly 10,000 individuals) whose biology and behavior will be measured using an unprecedented array of modalities over 20 years. It will also richly measure environmental conditions and events that KHP members experience using a geographic information system database of unparalleled scale, currently under construction in New York. In this manner, KHP will offer both synoptic and granular views of how human health and behavior coevolve over the life cycle and why they evolve differently for different people. In turn, we argue that this will allow for new discovery-based scientific approaches, rooted in big data analytics, to improving the health and quality of human life, particularly in urban contexts.

Key words: big data analytics; semistructured data; unstructured data

Introduction

There is ever-increasing evidence that from early in life our biology, the events we encounter, and the choices we make leave deep imprints on our minds and bodies that impact our future well-being, health, longevity—every aspect of our lives and our communities. Yet our scholarly understanding of this "bio-behavioral complex," this rich set of feedback effects between biology, behavior, and environment, remains surprisingly incomplete here at the beginning of the era of big data. As scientists working today, there is no escaping the fact that we lack some of the most basic longitudinal data about the bio-behavioral complex in domains ranging from education to finance to health. We have made radical advances ranging from the Human Genome Project, to the revolution in cognitive neuroscience, to the development of predictive psychological assays to innovations in social outcome measurement. But while our understanding of each of these subdomains has grown, we have made only incremental progress in uniting these many measurements in a manner that yields detailed behavioral phenotypes that characterize the myriad ways in which humans express their genetic endowment in different environmental settings.

This ignorance, with all its costs, is particularly surprising given two critical revolutions that have swept

[1]*New York University, New York, New York.*
[2]*The Kavli Foundation, Oxnard, California.*

*Address correspondence to: *Andrew Caplin, New York University, 19 W 4th Street, 6 FL, New York, NY, 10012*, E-mail: andrew.caplin@nyu.edu

across our academic and cultural landscapes: the development of massive discovery datasets in other scientific domains and the growth of the measurement technologies by which corporate big data has gained a deepening understanding of each of the isolated subdomains mentioned above. If one were to unite these many existing classes of available big data at the within-subject level, we believe that one could without a doubt produce a discovery dataset that would revolutionize the social and natural human sciences.

As an example of the role massive discovery datasets have played in recent scientific inquiry, consider the Sloan Digital Sky Survey. Until the 1990s, individual astronomers studied specific galaxies and quasars by booking time on established telescopes and searching the heavens for isolated data types relevant to their question at hand. In this way, astronomers laboriously aggregated small datasets ideally suited to resolving single hypotheses. In the late 1990s, however, the Sloan Foundation and its partners developed an automated telescopic system in New Mexico, the *Apache Point Telescope*, and began the robotic collection of a massive database that now catalogs photometric observations on over 500 million celestial objects across a huge range of data types. This kind of big data transformed galactic-level cosmology from a small data science to a big data science and has catalyzed a renaissance in astronomy and the initiation of many other astronomical catalogs of high scholarly impact. But despite the success of this big data approach with outward-pointing telescopes over the last decade, we have made no similar advances in our study of humanity with an inward-facing telescope.

One reason for this lapse in the study of humanity might be largely technical. Until very recently, we simply have not had the techniques and instruments required to build massive datasets at the scale and precision required to answer fundamental questions about the human condition. Over the course of the last decade, however, advances in computers, smartphones, the Internet, and large-scale biological measurement have made it possible to construct automated counterparts to the Sloan Apache Point Telescope for the study of humanity. In fact, isolated proprietary databases of this kind are now becoming commonplace. For example, Google regularly tracks the geolocations of hundreds of millions of people, credit-reporting companies track financial data about individuals to the level of individual purchases, and health insurance companies track medical and health related data at a similar granularity. Oddly though, no group has attempted to aggregate these datasets at the within-subject level in an effort to produce a Sloan Digital Sky Survey for Humanity.

In this article and the four that follow, we pose a simple question driven by these twin revolutions, the rise of truly massive discovery datasets in the physical and the natural sciences and the development of unconnected datasets on human health and behavior: What would be the advantage of generating a truly comprehensive longitudinal dataset that captured nearly all aspects of a representative human population's biology, behavior, and environment? In the pages that follow we argue not only that the aggregation of such a dataset is now possible, but also that it would provide fundamental advances in a host of bio-behavioral areas that could revolutionize scholarship and policy.

To make the potential of such a dataset clear, we first turn to four exemplar domains. We describe these four areas briefly in this article, as they serve as the detailed subjects of the four articles following this article. Our intention is to demonstrate the pressing need for such a discovery dataset for the big data community with four of many possible examples. We use these exemplars (and others that we have studied, which are not presented in detail here) to begin to identify the critical features required of a massive discovery dataset of this type. Finally, we describe a project now underway to launch the collection of just such a massive dataset in an urban center in the United States by the Kavli HUMAN Project (KHP)—a bio-behavioral counterpart to the Sloan Digital Sky Survey. The KHP is now being developed by an interdisciplinary research consortium and is designed to deepen behavioral phenotyping through enriched measurement and analysis. We discuss some of the technical aspects of engaging in such a large and long-term study in relation to the data storage technology, and how disparate data sources would be obtained, integrated, and verified. We stress flexibility of design to enable new data types to be pulled into KHP as time goes by—for example, as new and more effective activity or in-home monitors come on the market. To close the article, we revisit some of the exemplar domains, with KHP data in mind, to illustrate the practical importance of the project.

Why Big Data for the HUMAN Condition?

The case studies in this section, relating to aging, diet, smoking, and healthcare delivery, seem to us to indicate that interactions among biology, behavior, and

the environment are complex and dynamic even at the level of an individual human being. Four specific teams of KHP-associated experts, whose work is published in this issue of *Big Data* as companion pieces, have focused on the need for deeper bio-behavioral measurement. They are Dennis Ausiello,[a] Laura Bierut,[b] David Cesarini,[c] David Cutler,[d] Adam Drewnowski,[e] Ichiro Kawachi,[f] Kenneth Langa,[g] and Scott Lipnick.[h] The examples they present in these companion articles illustrate that behavior feeds back onto biological systems in myriad ways, calling for real-time tracking of both. They illustrate also that the broader environment we inhabit richly constrains and influences our behaviors and hence also our biological systems. Figure 1 illustrates key linkages between biology, behavior, and environment. The solid arrows going clockwise indicate directions of causation that are relatively well studied: from biology to behavior and from behavior to environment. The dashed arrows indicate two understudied directions of causation that motivate the KHP: feedback effects from behavior to biology and from environment to behavior. Note that these four arrows imply also that there are behaviorally mediated linkages between biology and environment.

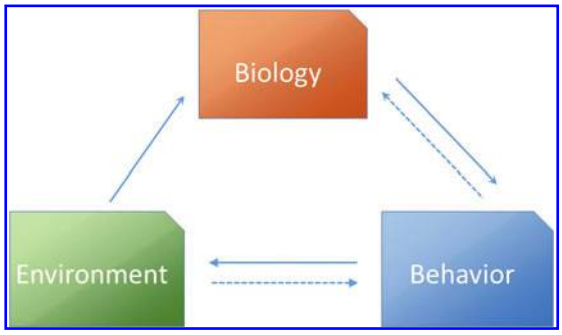


FIG. 1. The Biobehavioral Complex.

Case 1: Aging and cognitive decline

Aging and cognitive decline are massively important yet poorly understood. In a now-well-known study that appeared in the *Proceedings of the National Academy of Sciences* this year,[1] Belsky and colleagues identified a sample of 38-year-old Americans who came from a relatively homogenous sample group. They examined for each participant 10 biomarkers from the U.S. National Health and Nutrition Survey's (NHANES) data group. They found the "biological age" of these chronologically 38-year-old participants to range from 28 to 61, and to be approximately normally distributed with a standard deviation of more than 3 years. Further, they found that these differences in biological age were mirrored in differences in functional status, brain health, self-awareness of their own physical well-being, and facial appearance. Why? What causally accounts for the fact that some 38-year-olds function as 28-year-olds, while others function as 61-year-olds? What is the vector of characteristics that control aging?

Langa and Cutler (this issue[2]) point out that observed radically differential aging patterns are consistent with recent models that suggest that cognitive decline is accelerated by biological and social events throughout the life cycle. Yet the impact of behavior and environment on the process of aging and cognitive decline is understood in only a very general way, leaving the really important questions unanswered. For example, while correlations have been found between retirement and cognitive decline,[3] it is not known whether this is due to retirees' lower levels of mental engagement, reverse causation whereby cognitive decline induces retirement, or the resulting reduction in social contact. This is an issue we return to after introducing KHP.

The inability to resolve issues of causation reflects, to put it bluntly, a data limitation. As Cutler and Langa highlight in their article, cutting-edge questions are unanswered because we lack the data related to the

[a]Jackson Distinguished Prof. of Clinical Medicine, Harvard University; Director, Harvard Medical School MD/PhD Program; Emeritus Physician-in-Chief, Harvard Medical School; Member, IOM and AAAS.
[b]Alumni Endowed Prof. of Psychiatry and Co-Director of Outpatient Clinic, Washington University St. Louis, School of Medicine; Member, NIDA Genetics Consortium; Lead, Collaborative Genetic Study of Nicotine Dependence.
[c]Asst. Prof. of Economics, New York University; Center for Experimental Social Science; Co-Director, Social Science Genetic Association Consortium.
[d]Otto Eckstein Professor of Applied Economics, Harvard University; Research Associate, NBER; Council of Economic Advisers and National Economic Council, Bill Clinton Administration; Presidential Campaign Advisor to Bill Bradley, John Kerry, and Barack Obama; Senior Healthcare Advisor, Barak Obama Presidential Campaign.
[e]Prof. of Epidemiology, University of Washington, Seattle; Director of Nutritional Sciences Program, Center for Public Health Nutrition, and Center for Obesity Research, Univ. of Washington; Public Trustee, International Life Sciences Institute; Inventor, Nutrient Rich Foods Index and Affordable Nutrition Index.
[f]John L. Loeb & Frances Lehman Loeb Prof. of Social Epidemiology, Chair of Dept. of Social & Behavioral Sciences, Harvard University; Co-Director, Robert Wood Johnson Foundational Health & Society Scholars; Chair, Harvard School of Public Health's Institutional Review Board; Member, IOM.
[g]Prof. of Medicine, UMich; Research Scientist, UMich Veterans Affairs HSR&D Center for Clinical Management Research; Assoc. Director, Institute of Gerontology; Member, American Society for Clinical Investigation; Member, Health and Retirement Survey.
[h]Scientific Director of the Center for Assessment Technology and Continuous Health (CATCH) at Massachusetts General Hospital.

genetic regulators of aging processes; the impact of intrauterine growth restrictions and child maltreatment; the interaction of aging with cognitive stimulation in early, mid, and later life; the interaction of stress and physical activity; and the interaction of all of these with economic status. As Belsky et al.[1] stress, there is currently no data set suited to gaining an actual understanding of which factors contribute, and in what way, to aging in this sense (p. 6): "Our findings suggest that future studies of aging incorporate longitudinal repeated measures of biomarkers to track change. They also suggest that these studies incorporate *multiple biomarkers to track change across different organ systems* [our italics]."

What Langa and Cutler call for in their article is new synoptic measurement in the arena of aging and cognitive decline. In addition to the need for measuring biological factors going far beyond the admittedly groundbreaking NHANES measurements, there is a pressing need for the application of cognitive screening batteries at regular intervals as well as daily measurement of aspects of cognitive function with smartphone apps and other monitoring devices. There is a need for geolocation data to gauge how daily "life-space" changes as cognition declines, and there is even a need to conduct large-scale full brain imaging at key times in the lives of participants. All of these are critical if we are to assess the impact of cognitive decline on the ability of participants to perform key activities of daily living, to assess the amount of time that family members spend providing daily care to older adults with dementia, to assess the dynamics of the division of caregiving duties among family members, and to understand how these variables affect the work and family life of caregivers.

Case 2: Dietary choices and health

As with aging, the importance of diet to health and well-being is becoming increasingly clear. Indeed diet and longevity appear to be connected both theoretically and in practice. In their recent study, Belsky and colleagues[1] stress that better measures of aging should also be connected with improved measurement of diet to test for "the effectiveness of antiaging therapies (e.g., caloric restriction) without waiting for participants to complete their lifespans." Recent evidence suggests the potentially powerful impact of diet not only on lifespan but also on such diseases such as cancer.[4]

Unfortunately, actual large-scale measurements of diet, fully integrated with large-scale measures of biology and health at the within-subject level, remain unavailable. As Drewnowski and Kawachi (this issue[5]) point out, we know that obesity and some of its attendant ills are impacted by the decision to eat the refined grains, sugars, and fats that are energy dense, inexpensive, culturally appropriate, and widely accessible in our food supply. But as with aging, the limited scope and credibility of detailed and synoptic data on dietary choices over time prevents us from understanding the precise, and likely interdependent, roles that biology, economics, and psychology play in determining those food. While budgets and financial resources doubtless play major roles, it appears that some individuals and some economically disadvantaged groups are able to eat well for less. Other unknowns that hobble our understanding include such basic questions as the extent to which low-income urban residents shop locally for their food and the actual (rather than the assumed) importance of neighborhood-level accessibility to nutritious foods.

While the growth in diet-related health problems has induced a burst of research in the economic, epidemiological, and medical communities, lack of appropriate data is profoundly constraining our ability to make further progress. Cutler et al.[6] argue that there has been a significant increase in caloric intake possibly as a result of increased farm productivity and the increasing availability of highly caloric food and drink in convenient locations. Supporting data for this conclusion derive from measures of the food supply. Yet recent research suggests that, if the long-running UK National Food Survey Family Expenditure Survey is to be believed, caloric intake has in fact declined over time. It may therefore be that the increase in obesity in the United Kingdom and possibly even the United States is better seen as resulting from decreased activity rather than increased consumption. Of course, the key issue here is whether or not food diaries are credible, a question on which the jury remains out.

To understand the forces that affect food choice and how these in turn impact biological factors will require rich dietary measures and other social and behavioral measures. An important precursor in this regard is the Seattle Obesity Study (SOS), which has already provided first insights into the economic and geographic factors underlying food choice.[7] What marked the SOS as revolutionary was that, to improve measurement, perceived expenditures were validated using actual expenditures backed by two-week receipts for all foods purchased at home and away from home. But

to take the next step, it will be critical to measure the interactions between diet, economics, geolocations, neighborhoods, and biology over a prolonged period with far increased granularity and resolution. Even the impact of behavioral efforts to change eating habits has yet to be measured in depth. The difficulties that many individuals have in sticking with diet plans are unexplained yet may give broader insights into how brain, body, and behavior underlie self-damaging behaviors. Understanding how diet interacts with economic and social status, weight, aging measures, health, disease, and exposures to stress is possible—if we have the right data at the right scale.

Case 3: Smoking and health

Cigarette smoking is, in many ways, the modern poster child for a self-damaging addictive behavior. The basic biology of smoking is now well understood, with apparent roots in nicotinic absorption and associated dopaminergic responses. Recent studies have identified a single nucleotide polymorphism, rs1051730, colloquially known as "Mr. Big," which has been found both to alter the responsiveness of nicotinic receptors and to systematically impact measured smoking.

Given the prolonged scientific, medical, and social-awareness focus on smoking over the last several decades, one might expect smoking behaviors to be well-measured and characterized with sufficient granularity to allow us to comprehensively define the risk factors and treatment tools necessary for the mediation of smoking's impact on our societies. Unfortunately, this is far from the case, and this is profoundly limiting our ability to develop appropriate policy measures. Current survey-based measures of assessing smoking behavior are subject to recall biases and social stigma as well as limited granularity, properties that limit the utility of these data.

A stark demonstration of the limits implied by current survey methods is the work of Benjamin et al.,[8] who explored the impact of rs1051730 on measured smoking and on smoking-related disease in the recently genotyped Health and Retirement Study (HRS), a nationally representative longitudinal survey of Americans over 50 years of age that has set the standard for large-scale synoptic studies of the bio-behavioral complex. They find that, among smokers, those with two copies of the smoking-associated allele at rs1051730 had maximum lifetime smoking only roughly 10% higher than those with no copies. However, the effect on lung conditions "such as bronchitis or emphysema" is dramatically larger than that 10% would imply. Those with two copies of the dangerous allele are *30% more likely* to be diagnosed with these conditions than those with no copies.

How can a gene that has a modest effect on measured tobacco smoke intake have such a large effect on smoking-related lung disease? While more data are needed to pin down the channel of causation, a possible explanation for this asymmetry is that "smoking behavior" is only fragmentally measured in the HRS and other studies like it, and that a far stronger linkage would be identified with more comprehensive measures of smoking decisions over the life cycle. How many relationships like this may exist in arena of smoking (and other health behaviors) is a wide open question. Absent radically improved comprehensive studies and measurements, however, we have no way to discover such relationships or advance our overall understanding of these bio-behavioral-environmental complexes. Again, we revisit this issue after introducing KHP.

To overcome existing measurement limitations requires far more accurate real-time tracking of smoking behavior. Detailed measurements of purchasing behavior together with geo-tracks of subject locations (indicating, e.g., when they leave the workplace to stand still outside and smoke) and data from activity monitors would provide particularly high-resolution data on smoking behavior. Self-reported smoking quantities could be cross-checked against credit card records on cigarette purchases to yield within-subject calibration tools to better measure smoking rates. Biomarkers such as cotinine could also be measured in hair samples for longer-term bioassessments.

Given the clear evidence of smoking's feedback effects on biological factors, methods for improved measurements become even more critical. A robust epigenetic finding is that smoking is associated with the methylation of many genes. Methylation refers to the state of a DNA molecule. It is typically measured using methylation arrays that probe a certain number of genomic regions and, for each region, provide a numerical measure of the degree of methylation (a number between 0 and 1). Methylation is interesting to measure because it is an important mechanism for gene regulation, impacting how genes are expressed and proteins produced.[9] Hence, if certain genes are differentially methylated in smokers and nonsmokers, these differences may provide clues about the biological pathways through which smoking impacts health. Whether or not these methylation patterns can help

explain some of the biological pathways through which smoking ultimately impacts lung health and lung cancer is a vibrant area of research. But only by capturing both biology and behavior in a more precise and dynamic fashion can we resolve the ultimate linkages between smoking behavior and health.

Case 4: Bio-behavioral measurement and healthcare

As the above examples hint, and as further stressed by Ausiello and Lipnick (this issue[10]), we are in the throes of a major revolution in biological understanding. In addition to the recent and ongoing upheavals in our understandings of both genetics and neurobiology, the microbiome has emerged as a central stage for interdisciplinary biological research, with exciting breakthroughs already made and a vast uncharted territory yet to be explored. Microbes (bacterial microorganisms) colonize the gut at birth. In a striking example of bio-behavioral-environmental interaction, recent studies are suggesting profound impacts of behavioral and environmental factors on what types of microbes are present in our gut. In turn, this microbial aggregate appears to have significant influence on metabolism, immunity, and even behavior, in some animal models.

There have also been major advances in our understanding of inflammatory pathways. For example, a wide variety of inflammatory cells and pathways are being studied in auto-inflammatory disease as well as common diseases such as type 2 diabetes and cardiovascular disease. Biomarkers such as C-reactive protein or the erythrocyte sedimentation rate are nonspecific markers of inflammation currently used in clinical practice. A variety of new approaches could enable scientists to parse inflammation more precisely (including serum levels of specific cytokines or mediators), and create better assays for testing the activity of inflammatory cells (including molecular imaging of inflammatory cells, or microfluidic devices that can trap or analyze single cells).

In sharp contrast to the rapid and ongoing revolution in the biological sciences, translation of these discoveries into medical practices and applications has been taking place at a far slower pace. A key goal of the big data community should be to help bridge the gap between scientific advance and clinical practice. In this respect, an ongoing effort taking place at Massachusetts General Hospital is of particular importance for defining what measurements need to be incorporated in future synoptic studies. The newly formed Center for Assessment Technology and Continuous Health (CATCH) provides an important model for the collection of comprehensive phenotypic data of this kind as part of more synoptic efforts. To capitalize and expand upon the ideas and lessons coming out of CATCH requires a study that drills down in greater depth into the behavioral patterns and environmental exposures that interact with health outcomes. By enabling more facile and passive quantification of environmental exposures, such a study would create an important new data resource that can be integrated with genetic, clinical, and behavioral information and thereby enhance our understanding of the complex forces that shape human health.

What Should the First Synoptic Study of Humanity Include?

Until now, large-scale longitudinal studies have generally been focused on specific domains of inquiry or subsets of the population. They provide detailed catalogs of genetics or health records or data about finances, or even more integrated data about health and finances, but do not examine the complete dynamic pattern of human behavior, biology, and environment across the lifespan in a single group of subjects.

The main exception to this rule is the U.S. HRS, a nationally representative longitudinal survey of Americans over 50 years of age and their spouses. The initial HRS sample was collected in 1992 and more cohorts have been added over time. Most of the recent advances in our understanding of cognitive decline, smoking, and other areas of linkage between biology and behavior derive from HRS data. So successful is it that variants have been developed worldwide in dozens of countries. The importance of these surveys stems both from their representative nature and from their breadth of coverage, including as they do genetic, cognitive, health, financial, psychological, and demographic factors.

For all its great value, however, one key limitation of the HRS is its age restriction. A second is that it is administered every two years with only minor adjustments from wave-to-wave and really quite limited data are gathered at each two-year sample. Moreover, biological measures are generally made only once on each subject. A final limitation is that coverage is based on the individual rather than the family unit and community. What we believe is now needed is an HRS for the big data revolution that removes these constraints and builds fundamentally on the longitudinal survey revolution initiated by the HRS.

Diversity of data sources

A truly synoptic study would have to gather information from study participants across numerous domains, as listed in Table 1, in order to provide a 360° view of the study participants. Information gathering would make use of a variety of methods, some of which would involve study participants directly, such as gathering blood samples for full genome sequencing (3 billion base pairs), in-person tests to assess participants' psychological well-being and cognitive status, and use of smartphone apps to gather geo-location data at regular intervals. Similarly, the study would seek participants' authorization in order to receive copies of financial records, such as tax filings, monthly statements for bank accounts, and credit cards. Lastly, the study would utilize public databases that are maintained by NYC and NYS governments, such as those on census data, education, and crime statistics.

Timeline

In order to provide longitudinal data that maximize the power of the study, it is essential that any such project has a relatively long time horizon. At a minimum, a five-year horizon would allow one to begin to realize the potential of such an undertaking. But in keeping with the goal of building a truly revolutionary resource, we believe that a 20-year study duration would be required to capture the long-term impact of environmental features like education, chemical exposure, and the human lifecycle.

Sample group

Many recent large-scale studies have focused on gathering data from "samples of opportunity," groups of participants selected because they have a particular disease (or set of diseases) or are members of a particular social group. Far more powerful would be the generation of a statistically representative sample that allowed the study to capture data about our society rather than about a subgroup. While building a representative cohort is more complicated and more expensive, we believe that it would be essential if the undertaking were to achieve its potential. The minimum size for such a study would be approximately 10,000 individuals from approximately 2,500 family units. We believe that this represents a reasonable compromise between the desire for a large sample for reasons of statistical and social power and the high costs of subject recruitment, retention, and bio-behavioral measurement.

Sample quality and volume

The study would have to ensure the highest attainable sample quality and volume by employing several key measures. The corner stone of the data strategy would have to be to recruit a sufficiently large and representative sample of subjects. To do that, the study could begin by focusing on a single urban area where data of particularly high quality are already available at low cost. We believe that an ideal initial venue would be a cross-sectional, demographically representative assessment of residents of the city of New York. (Although data of sufficient quality should soon be available in several other urban centers in the United States.) It is worth noting that many large-scale U.S. studies fail over this issue.

In addition, one needs to approach data quality with close consideration of the specifics of each information domain. For instance, to collect data about participants' movements and activities at a reasonable cost, one needs to leverage mature and widely used data collection platforms, such as smartphones (iOS and Android) and activity trackers, while developing custom "apps" designed around the needs of the synoptic study. In collecting biomedical and psychological data, the study group must work with highly qualified agencies and scholars to ensure repeatable and reliable test results. Lastly, one absolutely must establish proactive monitoring of data inflow in order to identify and resolve potential issues before they can impact the overall quality of data. Such issues may be due to factors such as participants' failure to follow study requirements, or technical problems. While these are fundamental tenets of big data today, they absolutely must be incorporated as fundamental features of any synoptic study of this kind.

Such a project must, we believe, capitalize on the current technologies that enable the rapid acquisition of substantial amounts of electronic data. Automated data collection enhances the comprehensiveness of the available data, but it can also help with verification of data as it provides multiple views of any particular event. For example, location data for an individual should coincide with purchase information at a store at that same location. Of course, there will be particular areas where automated data collection will be difficult, if not impossible, and we will have to rely on proxies or more granular level data when these problems arise. While it is unlikely that people will be willing to keep detailed food diaries over the course of the study, detailed purchase information from restaurants,

Table 1. Domains of data collection from KHP study participants

Information domain	*Data sources*	*Inputs into KHP study*
Demographics	Participant questionnaire Supporting documentation, e.g., birth certificate, driver's license, or passport	Demographic information about participant household and individual members of the household, such as age, gender, and ethnicity
Home environment	Participant questionnaire Building information Survey and measurements by KHP field team Sensors for air quality and ambient noise Utility records	Information about housing space, presence of toxins, air quality, ambient noise level, and water and energy use
Neighborhood baseline	NYC public data sets on census, education, law enforcement, public service, and GIS NYU CUSP databases	Information about the neighborhood in which the participant lives, such as demographic composition, median income, school ratings, emergency service requests, and crime statistics
Biomedical	Physical exam (weight, height, BMI, resting heart rate, blood pressure) Blood sample (for genetics and blood chemistry) Urine sample (for toxicology) Saliva sample (for oral microbiome, genetics and stress measurement) Hair sample (for toxicology and chemical exposure) Stool sample (for gut microbiome) Electronic medical records (EMRs), doctor's notes, dentist records, and hospitalization history Health insurance records NYS database on prescriptions (SPARCS) Silicone wristbands (for chemical exposure) In limited numbers: functional MRI, electroencephalogram, and electrocardiogram for more invasive study of core set of participants	Information about each participant's medical and dental history, physiology, biochemistry, complete whole genome genetics, complete microbiomes, and complete pharmacological use profiles
Diet and health	Participant food diaries (for limited duration, repeated regularly) Financial transaction records, mined for food- and health-related purchases	Information about each participant's diet, use of alcohol, tobacco, and other substances
Psychological	Structured interviews of participants by trained professionals Self-administered tests on smartphones and tablets	Information about participants' mental health, personality attributes, levels of cognitive function, executive function and memory, and risk preferences
Educational	Participants' educational records and extracurricular activity records Survey of participants' homes by KHP field team NYC Department of Education databases on school rankings and progress of individual students	Information about participants' formal and informal educational history (e.g., number of books in the home) and progress of current education
Occupational	Participants' curriculum vitae (oral or written) Participants' W-2 records	Information about participants' occupational history and progress of their occupation/career during the study time frame.
Activity	Smartphone app (for location, activity, and socializing data) Wearable trackers Bluetooth-based presence sensors in participants' home Smartphone/tablet app for social media and digital contacts NYC GIS database	Information about the times and duration of different activities, such as sleep, commute/travel, work/school, exercise, entertainment, socializing, and screen time, as measured by wearable technologies, smartphone apps, and presence detection systems
Family interactions	Participant questionnaire Bluetooth-based presence sensors in participants' homes	Information about the frequency and duration of interaction between parents and children in the home Information about the level of care given to family members by family members
Financial	Participant questionnaire W-2's Title and ownership documents for key assets Bank records Credit card and debit card records Loan records Public assistance records (e.g., SNAP) Retirement planning account information (e.g., pension, 401k) Rental agreements, mortgage records	Information about participants' sources of income, major assets and liabilities, categories of expenses, savings, and retirement planning activities. Detailed purchase data to the level of all individual purchases, grocery purchases at the level of individual items, prescription drug co-pay data, alcohol purchases, tobacco purchases, etc.
Interactions with law enforcement	Participants' call history NYC Police Department databases on 911 calls, 311 calls, stop and frisk activity, and arrests NYC District Attorney databases on case histories	Information about participants' interaction with law enforcement agencies as either victims or potential culprits

high-frequency photographic records, and grocery store receipts can still provide meaningful data about diet that transcend these kinds of problems. Cash purchases are another area in which indirect measures are required to make inferences. However, in all cases, the extraordinarily detailed and multifaceted data collection possible with current electronic technologies offers the opportunity to estimate proxy measures for economic and behavioral factors that have not previously been studied quantitatively at scale by any group.

One of the key objectives of any such study must be to identify early predictors of health outcomes, even before such outcomes can be diagnosed, such as identifying subtle changes in behavior that may indicate eventual onset of a disease (e.g., Alzheimer's or Parkinson's). A key determinant of data quality for a time-series analysis of this kind is sampling frequency, where high sampling frequency enables researchers to observe small or transient changes that may ultimately turn out to be reliable predictors. High-frequency data collection is thus critical, the frequency depending on the expected rate of change in each information domain, if we are to obtain a "high-definition" view of study participants' lives.

It would also be crucial to gather detailed historical information on medical and environmental experiences. Data collection of this kind should include complete genetics; complete microbiomes; standard physical examinations (including lab work and psychological examinations); social networks/safety nets; geolocation data; activity tracking (by band or by phone); sleep tracking; hair analysis, urine analysis, and pharmacy records for assessing drug use; direct measures of stress levels; environmental quality (air, noise, chemical exposure via the standard silicone wristband approach); detailed purchasing data (particularly food types), work experience (most Americans spend as much time at work as at home, and the effects of heavy physical demands, a sedentary setting, or chronic stress are likely to have substantial influences on health outcomes); detailed structured and unstructured medical records (including ongoing examinations and treatments); and social services records (disability, Medicaid, Medicare, SNAP).

Ease of Integration with Third-Party Data Sources: Advantages of New York City for Such an Initial Study

A number of active projects by New York City offer the opportunity for added power in a synoptic study that could not be achieved elsewhere in the world at this time. One of these is a recent project by the NYC Health and Hospitals Corporation (HHC) to modernize the electronic health records system. It is expected that, by 2017, electronic medical records from across all NYC HHC patient care facilities, including hospitals, long-term care facilities, diagnostic treatment centers, and community-based clinics will be fully integrated. The NYC HHC is the largest municipal health system in the country, and treats about 1.4 million patients a year, including a large proportion of the uninsured. The HHC medical record system will be conjoined with a consortia of New York's other large-scale medical providers, which should yield the most comprehensive electronic medical record system in the United States. Access to that database provided by conducting the study inside New York City would make the proposed dataset of exceptional use to medical researchers.

Another data resource that would be available to a New York–based study is the Statewide Planning and Research Cooperative System (SPARCS). The SPARCS database contains individual-level detail on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient visit (it includes ICD-9 codes and data on ambulatory surgery, emergency department, and outpatient services), and each ambulatory surgery and outpatient services visit to a hospital extension clinic and diagnostic and treatment center licensed to provide ambulatory surgery services. In addition, New York City maintains a relatively new prescription-drug monitoring program registry for all controlled substances, which contains individual-level records.

Another significant external data source available only in New York at this time (although Chicago is rapidly also developing such a system) will be the Geographic Information System (GIS) that is currently being built by New York University as a resource for the study of New York. This database will provide a multilayered view of New York City, capturing information about air pollution, electricity use, garbage volume, emergency service requests, noise complaints, and even parking tickets issued and the individuals to whom they are issued. Overlaying this information with data that are collected from study participants would enable investigations of interactions between the environment, behavior, and biology to understand how these factors turn a predisposition for something like heart disease, diabetes, or depression into pathology in only a subset of those who have the vulnerability.

Privacy and security

Given the extensive view into study participants' lives that such a project would provide, getting privacy and security aspects of the study "right" would have to be one of the study's highest priorities. On the privacy front, any such study would have to provide clear, concise, yet comprehensive language to obtain the necessary authorizations from study participants, and it would have to filter any personally identifiable information from the data set made available to researchers. In addition, it would have to use federated data storage with multilevel state-of-the-art authorization and security protocols to prevent participants from being re-identified through the usual, customary, and reasonable manipulation of the data set. Lastly, the standard output of the data for any research activity would have to be aggregate analysis, without the specific underlying records at the individual and/or event level.

On the security front, any such study would have to use strong encryption for data storage and it would certainly not provide direct access to the core data set from public networks. It would also have to employ strong security monitoring procedures and tools to detect and stop unauthorized access to any part of the data with great expediency. For each research effort that will utilize any of its data, the study should generate a data "slice" in the form of a so-called "data mart," which would include only data fields that are relevant to the study, rather than providing any specific researcher access to the entire data set. Access to each data mart would have to be heavily monitored and each data mart should be deleted in its entirety at the conclusion of the relevant research effort.

Data processing platform

The project's choices for data processing platforms would have to be forward-looking in order to adapt to medical and technological advances during the lifetime of the study, which will undoubtedly introduce more accurate measurements of existing data over time, for instance, through the use of more accurate air quality sensors. It is also reasonable to expect that new data types would enter the study, such as blood glucose level measurements from temporary tattoos. At times, new data points would also become available retroactively, for instance, when a new blood test would be applied to stored blood samples that had been collected years earlier. In addition, as existing NYC and Center for Urban Science and Progress (CUSP) databases expand their capabilities, the study would have the opportunity to incorporate or link to increasingly more granular third-party data. While the specifics of future data types and impacts to study population cannot be known at the outset of the project, we can make two projections with high confidence: (1) as the project progressed there would be a broader variety of data points, both from study participants and from external sources, very possibly collected at higher frequency, and (2) as a result, data volume and processing needs of the project would also continue to grow.

Data architecture for the project would have to plan for these changes in its design and selection of database platform. This architecture would have to incorporate flexibility to define new data points in the future without requiring a major overhaul of the database, while not compromising on performance a great deal. Similarly, metadata about collection methodologies would need to be incorporated into the data architecture, for instance, makes and models of sensors, their measurement sensitivities, and the dates when those sensors were in use for each participant. Lastly, a flexible document database under "NoSQL" umbrella, for example, MongoDB, would very possibly be deployed in the study.

In order to support the expected growth in processing demands, the project would have to utilize scalable clustered solutions, such as Hadoop/MapReduce, and later evolutions of those technologies. All implementation decisions would also need to take into account the costs and benefits of creating an "enterprise" solution versus utilizing cloud-based solutions, such as Amazon Elastic MapReduce for a Hadoop platform or Amazon RedShift for data warehousing, either end to end or for specific functional needs. In these decisions, the project would seek state-of-the-art solutions that offer proven security, reliability, and performance.

It is worth noting that as we design the project we are very mindful of the "janitor" problem in big data, which estimates that between 50% and 80% of the effort on big data projects is on "janitorial," or "data wrangling," tasks. This issue has been discussed in the popular press, as well as within the big data community. By paying painstaking attention to data quality at the point of collection, we hope to significantly reduce the scope of data wrangling efforts for future researchers.

Support for analytics

The project would have to support widely used analytics languages, such as *R* and *Python*, to support the data

science community. We believe that it should host a library of widely used analytics tools as a user community grows. It should also develop a basic set of data analysis and visualization tools for newcomers to big data, along with expert computer scientists working in-house to support analysis by scholars unfamiliar with big data approaches.

Data sources and data ingestion

The project would need to work with both unstructured and structured data from analog and digital sources. Unstructured analog data, for example, could take the form of hand-written notes by healthcare professionals. This data set would be scanned and stored for eventual use of automated handwriting analysis, except in a limited number of cases. Structured analog data would take the form of standardized test results (e.g., medical test results), which can be converted into structured digital data with ease using existing technologies. Digital structured data would include input from digital devices, such as smartphones, wristbands, and Bluetooth beacons, as well as data sourced from external databases, such as Electronic Medical Reports and NYC GIS. Figures 2 and 3 illustrate a proposed approach to data ingestion.

Record validation

Record validation would be an area of particular focus in the KHP to ensure that incoming data would be complete and accurate. Records in each data stream would be validated in two stages, taking into account all available information both from the particular data stream and other data streams.

The first stage of record validation would be point-wise validation, which would ensure that observed data for a particular data point (variable) arrived at the expected frequency and volume, in expected formats, and were within the valid range of values for the variable. For instance, geo-location data from participants' smartphones should include valid latitude and longitude values, and they should arrive every few minutes, taking into account cases when the smartphone may be without cellular or Wi-Fi coverage, or without power.

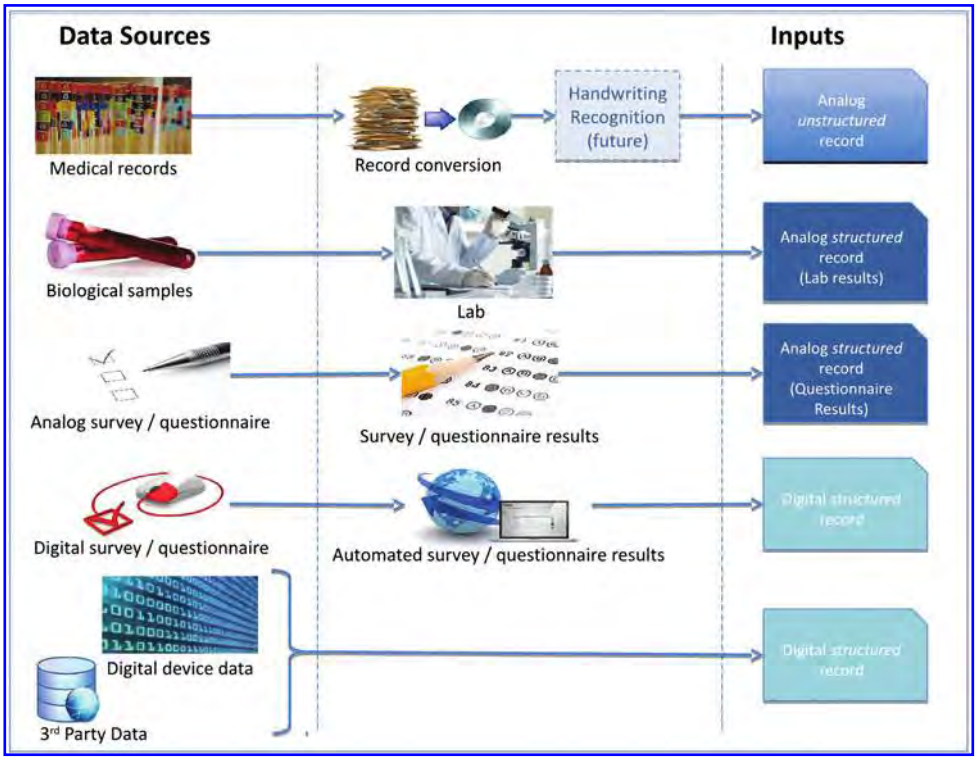


FIG. 2. Data sources to inputs for the KHP study.

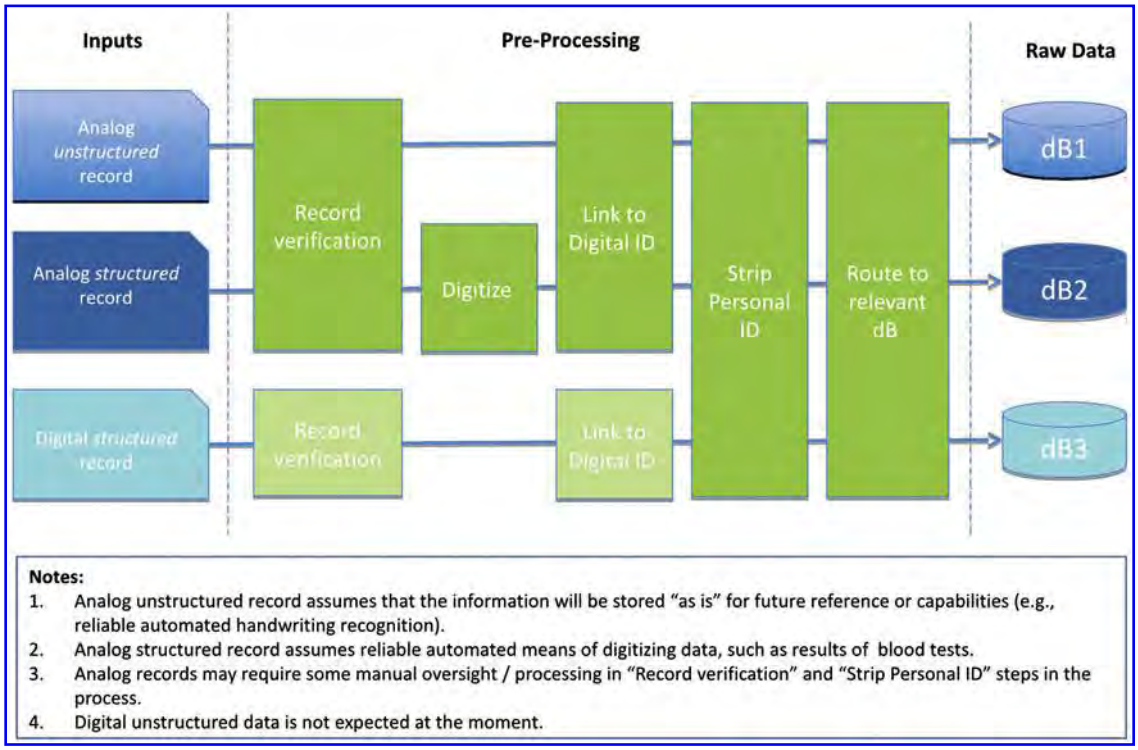


FIG. 3. Data ingestion for KHP study.

The second stage of record validation would ensure that each data point is accurate through three mechanisms: cross-validation, predictive checks, and error correction. The first mechanism, cross-validation, would leverage other data streams, when possible, to crosscheck individual records. For instance, when a participant was at home, presence data from Bluetooth-based sensors would have to be consistent with geo-location data from the participant's smartphone. The second mechanism, predictive checks, would use historical data for the particular data stream and from other relevant data streams to predict expected values for a data point, and compare the observed value to the predicted value, in order to highlight potential "exceptions." For instance, a short-term prediction for the expected geo-location of a participant could leverage most recent geo-location and velocity data to predict the participant's location within the next few minutes. A longer-term prediction using a longer-term behavioral history of the participant could identify transient or permanent shifts in participants' movement patterns. For example, a change in morning commute routine could indicate a change in employment, or it could be a transient change that is not repeated. The third mechanism, error correction, would utilize conservative error-correction measures to compensate for records with potential errors and for missing records. This mechanism would only update an erroneous or missing record when it has high confidence; otherwise, it would remove potential errors from input and leave missing records untouched. Error correction could be applied to all possible data streams. For example, it could be applied to a blood test in the following manner: when lipid values in a blood test did not align with the expected range of values, given the physiology and medical history of the participant, one possible error correction approach would be to order a second test in order to eliminate a potential error.

Availability to scholars

Of course, any such synoptic study must be widely available to scholars from any discipline. While preserving the privacy and security of participants will be essential, an open door policy for access to the data is required if the potential of the dataset is to be realized.

KHP, Big Science, and Big Data

It is clear that a new approach—a comprehensive view—is necessary to understand how the interactions between genetics, mind and body, behavior, and environment interact with human health. Until now, large-scale longitudinal studies have been focused on specific domains of inquiry or subsets of the population. As a result, existing large-scale datasets have provided detailed catalogs of genetics or health records or data about finances, or even more integrated data about health and finances, but no study has yet examined the complete "360-degree" dynamic pattern of human behavior, biology, and environment across the lifespan in a single group of subjects.

Over the last year, a team of researchers supported by the Kavli Foundation has completed the initial design phase for a large-scale study along the lines described in this article. We refer to this study as the Kavli HUMAN Project: Human Understanding Through Measurement and ANalytics. The data we plan to collect, beginning in early 2017, will follow the blueprint laid out in the previous section. It will be both broad and deep—many hundreds of terabytes of information per year about individuals, families, and the environment in which they live and work. It will include complete genetic sequencing, electronic medical records, psychological assessments, social network and communication pattern profiling, education data, employment data, financial data, and location for each participant.

Alongside this detailed catalog of information about each individual will be a multilayered database of New York City: information about electricity use, garbage volume, emergency service requests, noise complaints, and even parking tickets issued in the places where these individuals live, work, and play. As noted above, the KHP will take advantage of the unique resources available in New York City, the HHC electronic medical record initiative, and the SPARCS database, as described above. Another NYC-specific resource available to the KHP is the extraordinarily rich and detailed collection of the city's administrative and operational datasets. These document the urban milieu in which the study participants live and work. New York City has one of the largest collections of publicly available datasets in the United States, but as part of a partnership with New York University's CUSP, KHP will have access to an even wider range of data gathered around the city, including local chemical release plumes measured by hyperspectral sensors.

In order to facilitate the truly interdisciplinary scholarship that the KHP can enable, the members of the KHP team have highly diverse academic backgrounds, but share a deep commitment to interdisciplinary collaboration, and the conviction that improvements in measurement combined with robust theory are the keys to such progress. Key members of the central KHP organization are Project Leader Paul Glimcher (neuroscience, psychology, and economics); Directors Steven Koonin (physics and data science), Ari Patrinos (genetics and environmental science), Andrew Caplin (economics and data engineering), and Elizabeth Phelps (psychology); Chief Scientist Hannah Bayer (neuroeconomics); and, critically, as observer of the KHP, Miyoung Chun (biology), executive vice president of science programs at the Kavli Foundation, who has played a key leadership role in the effort to bring biological and social sciences together—joining mind to body, and mind–body to society.

The core leadership structure of the KHP is supported by five domain-specific academic boards, or advisory councils—each chaired by a leading scholar or practitioner in that domain. The Scientific Agenda Advisory Council identifies use cases for KHP data and publicizes them by publishing White Papers and is chaired by Andrew Caplin. The Study Frame Design Advisory Council defines the number of subjects and the composition of the core subject pool in order to provide enough power to address research questions about human behavior and is led by Kathleen McGarry. The Measurement and Technology Advisory Council is responsible for identifying the traditional and novel approaches that will be used to measure biology, behavior, and the environment, and is led by Alex "Sandy" Pentland. The Privacy and Security Advisory Council designs KHP's privacy and data ownership policies, and specifies the data security technologies necessary to ensure the safety of the data while preserving access to researchers, and is led by Lynn Goldstein. The Education and Public Outreach Advisory Council seeks to educate key constituencies, including the academy, the press, the public, and policy makers on the research and the findings, and it will be convened as the KHP moves closer to launching subject recruitment and data collection.

The KHP stands poised to capitalize on the recent expansions in electronic record keeping, new methods for the management and analysis of large datasets, and advances in stationary and mobile data collection that have dramatically changed the information technology

landscape. Large-scale information gathering is now possible at relatively low cost, offering a novel opportunity to go beyond previous explorations and to perform a much more extensive, much more thoroughly integrated survey and analysis of human behavior. This synoptic study of humanity will provide the opportunity to go beyond disciplinary boundaries and make substantial progress in understanding the dynamic interplay among environment, biology, and behavior.

The Scientific and Policy Importance of KHP

To illustrate the powerful insights that the KHP study may enable, we now revisit several of the cases introduced previously with the KHP data in mind. Consider first the unanswered question of why cognition appears to fall at retirement. As outlined above, a prominent hypothesis is that retirement leads to a reduction in mental activity, and that those that stop using their mind tend to lose it at a faster rate. A second hypothesis is that it is the reduction in social contact associated with retirement that hastens decline. A third is that it is due to reduced physical activity that is indirectly linked with mental decline. Finally, this may be a case of reverse causation, with those who decline early tending as a result to retire early.

This is a case in which the advantages of the KHP over existing data sets would be overwhelming. Reverse causation would be analyzed by tracking of physical and cognitive performance in the period before retirement. Social contact, physical activity, and mental activity would also be directly monitored. While one might expect the overall patterns that have been noted in the HRS and other data sets also to be present in KHP data, the differential time paths of physical, cognitive, and social forces would provide definitive evidence on the relative importance of the channels that have been theorized about. With this enriched scientific evidence, policies could be adopted to help provide those on the verge of retirement with information about palliative measures to guard against possible subsequent decline. The findings may also be of great social and policy importance given the rapid aging of the population and the private and social losses associated with cognitive decline. More accurate scientific knowledge might lead many to stay longer in the labor force and/or to choose occupations that produce the appropriate form of mental stimulation. In turn, employers would be incentivized to enrich the work environment to keep their employees engaged and productive for as long as possible.

Our second illustration of the value of the KHP relates to smoking behavior. As indicated above, it is now known that effects of genetic factors on self-reported smoking levels are swamped by their effects on lung health and death rates. Each measured cigarette appears to do more harm to those with genetic factors that expose them to high as opposed to low risk. This seems entirely baffling since there is no known biological basis that makes such a difference credible. The most likely hypothesis is that the flaw lies in the very poor measures of smoking that are used in genetic studies as well as other important bio-behavioral data sets such as the HRS.

As indicated above, the advantage of the KHP in terms of measurement of smoking behaviors is overwhelming. In essence, it will enable these to be tracked at very high resolution over long periods, together with detailed and extensive longitudinal health measures. There is then every reason to hope for a resolution of this scientific mystery. KHP data may reveal that the actual number of cigarettes smoked over the life cycle differs far more across genotypes than do existing crude measures of smoking. Alternatively, it may find that those with the less risky allele find it relatively easy to quit as their health starts to deteriorate, enabling them to arrest the damage that smoking does at a relatively early stage. Finally, it may be found that the genotype influences more sophisticated behaviors, such as depth of drag, length of time holding smoke in the lungs, or manner of smoking down cigarettes. The KHP will clearly help resolve the relative contributions of these different modulatory influences. In the process it will help to pinpoint palliative strategies, and clarify precisely those times at which early warning may be provided to those with riskier alleles.

In addition to highlighting the value of new data, the case of smoking illustrates the profound synergies between data sets that operate at different levels of depth and breadth. The original findings on smoking and genes resulted from combining data from genome-wide association studies (GWAS) with advances in biological knowledge of the process of nicotine absorption. Use of the HRS was then invaluable in connecting this with health outcomes. To go further requires KHP data that is capable of separating explanatory hypotheses by virtue of its far greater granularity. There is every reason to expect massive numbers of new GWAS findings to appear over the years with ever richer understanding of the underlying cellular mechanisms. Many of the most important findings are likely to

replicate in the HRS, and point to key interactions with health, wealth, and other important outcomes. Having a more granular data set such as KHP will then prove invaluable in advancing our understanding of the precise channels of effect and appropriate policy responses.

Our final use case relates again to the broad area of aging and cognitive decline. The study by Belsky et al.[1] involves many biomarkers indicative of biological age. Yet the authors acknowledge that the set of biomarkers that can provide the optimal prediction of age-related phenomena is not yet known. They also acknowledge several other study limitations:

- Analysis of a single cohort lacking ethnic minority populations.
- Data were collected only three times, once every 6 years, for a single birth cohort.
- Lack of repeat measurements of biomarkers that might better quantify aging.
  The KHP would address these limitations through its current design.
- The study population would be a cross section of New York City residents, including many ethnic minorities.
- The study population would provide biological samples every three years, which would deliver repeated measurements and allow more accurate tracking of changes over time.
- The study population would include all age groups, both younger and older individuals, over a time period of 10–20 years, providing information on how specific biomarkers operate at different life stages.

In addition, when multiple factors impact the rate of change of a biomarker, the KHP would enable researchers to identify which factors may contribute to changes in biomarkers through its extensive data set. For instance, through the KHP study data set it would be possible to determine what environmental, psychological, or behavioral factors correlate best with an observed change in HbA1c (glycated hemoglobin), which is a biomarker for diabetes. In such an investigation, it would be possible to determine whether HbA1c levels are more sensitive to physical activity, calorie intake, nutrition quality, or hitherto unsuspected factors.

Lastly, as the KHP would focus on family units, it would also provide insights into what genetic information is modified from one generation to the next, and how fast each successive generation ages and experiences cognitive decline.

The above case studies are but a few exemplars of the potential for the KHP to revolutionize our understanding of bio-behavioral interactions and our ability to implement these advances in policies that improve the quality of life. A number of other case studies are currently under development as White Papers, and these indicate the breadth of areas in which the KHP can make significant contributions to the advancement of science and policy. In neuroscience, Wolfram Schultz is contributing on the balance between reward signals and self-control problems and Russell Poldrack on neuroeconomic measurement. Steven Koonin is writing on energy usage and Ari Patrinos on broader environmental issues. Charles Manski and Pamela Guistinelli are contributing on secondary education; B.J. Casey and Catherine Hartley on the forces that shape adolescent brain and behavior; and Regina Sullivan on child abuse. At the opposite end of the lifespan spectrum, Andrew Caplin and Kathleen McGarry are contributing on long-term care. Robert Townsend is writing on household finances to ensure integrity of these measurements. At the macro level, Edward Glaeser is teaching us how to track the big picture issues of urban and social economics.

Conclusions

We believe that there are compelling reasons for the big data community to begin to explore the possibility of a truly synoptic overview of the human condition at a within-subject level. The growth of big data technologies and the falling cost of human-related data capture by corporate actors have opened the door to this possibility. Just as the Sloan Digital Sky Survey and the Human Genome Project revolutionized the disciplines of astronomy and genetics, a large-scale synoptic study of a population could revolutionize our understanding of human behavior, health, and well-being. It could answer age-old questions about the interaction of education, diet, poverty, development, and technology with all aspects of the human condition. The KHP is an effort to develop the first study of this type but it is obviously not the last such effort. We seek to define a way for big data to be organized to understand the human bio-behavioral complex and serve as a platform for future efforts to understand the human condition.

Author Disclosure Statement

No competing financial interests exist.

References

1. Belsky D, Caspi A, Houts R, et al. Quantification of biological aging in young adults. PNAS. 2015;112:30.
2. Langa K, Cutler D. Opportunities for new insights on the life-course risks and outcomes of cognitive decline in the Kavli HUMAN Project. Big Data. 2015;3:189–192.
3. Rohwedder S, Willis R. Mental retirement. J Econ Persp. 2010;24:119–138.
4. Brandhorst S, Choi IY, Wei M, et al. A periodic diet that mimics fasting promotes multi-system regeneration, enhanced cognitive performance, and healthspan. Cell Metab. 2015;22:86–99.
5. Drewnowski A, Kawachi I. Diets and health: How food decisions are shaped by biology, economics, geography, and social interactions. Big Data. 2015;3:193–197.
6. Cutler D, Glaeser E, Shapiro J. Why have Americans become more obese? J Econ Persp. 2003;17:93–118.
7. Drewnowski A, Moudon AV, Jiao J, et al. Food environment and socio-economic status influence obesity rates in Seattle and in Paris. Int J Obes (Lond) 2014;38:306–314.
8. Benjamin D, Caplin A, Cesarini D, et al. Smoking, genes, and health: Evidence from the health and retirement study. NBER 2015; http://cess.nyu.edu/caplin/wp-content/uploads/2015/09/smoking-genes-and-health.pdf.
9. Lim DHK, Maher ER. SAC review DNA methylation: A form of epigenetic control of gene expression. Obstet Gynaecol. 2010;12:37–42.
10. Ausiello D, Lipnick S. Real-time assessment of wellness and disease in daily life. Big Data. 2015;3:203–208.

Cite this article as: Azmak O, Bayer H, Caplin A, Chun M, Glimcher P, Koonin S, Patrinos A (2015) Using big data to understand the human condition: The Kavli Human Project. *Big Data* 3:3, 173–188, DOI: 10.1089/big.2015.0012.

Abbreviations Used

CATCH = Center for Assessment Technology and Continuous Health
CUSP = Center for Urban Science and Progress
GIS = Geographic Information System
GWAS = genome-wide association studies
HHC = Health and Hospitals Corporation
HRS = Health and Retirement Study
KHP = Kavli HUMAN Project
NHANES = U.S. National Health and Nutrition Survey
SOS = Seattle Obesity Study
SPARCS = Statewide Planning and Research Cooperative System

Big Data
Volume 3 Number 3, 2015
DOI: 10.1089/big.2015.0016

ORIGINAL ARTICLE

Real-Time Assessment of Wellness and Disease in Daily Life

Dennis Ausiello[1,*] and Scott Lipnick[1]

Abstract

The next frontier in medicine involves better quantifying human traits, known as "phenotypes." Biological markers have been directly associated with disease risks, but poor measurement of behaviors such as diet and exercise limits our understanding of preventive measures. By joining together an uncommonly wide range of disciplines and expertise, the Kavli HUMAN Project will advance measurement of behavioral phenotypes, as well as environmental factors that impact behavior. By following the same individuals over time, KHP will liberate new understanding of dynamic links between behavioral phenotypes, disease, and the broader environment. As KHP advances understanding of the bio-behavioral complex, it will seed new approaches to the diagnosis, prevention, and treatment of human disease.

Key words: big data analytics; big data architecture; big data industry standards

Introduction

The field of genetics has exploded since the mapping of the human genome, but despite the treasure trove of information it provides, the human genome alone does not liberate comprehensive understanding of the human condition. The next great leap in human measurement and analysis is mapping the phenome: the sum total of human traits. The Kavli Human Project (KHP) will greatly advance this process.

Advances in human genetics have identified many genetic contributors to disease risk. At the same time, there are ever more extremely large datasets available containing multiple types of data relevant to human health. Growing analytic and computing capabilities are enabling mining of these large datasets for insights at the level of individual patients or entire populations. Yet the methods used to diagnose disease have conspicuously lagged behind these recent exciting discoveries. This is because disease traits are also influenced by largely unmeasured environmental and behavioral factors.

The KHP will liberate progress by convening a wider range of expertise than is traditional, including device engineers, front-line physicians, geneticists, and experts in behavioral modeling. It will focus directly on the development of novel quantitative human measurements, or phenotypes. By tracking individuals in rich detail over a long period of time, the KHP will capture and catalog dynamic patterns of individual and social behavior in new and richer metrics than ever before. This will help us develop new approaches to measuring human health, and so we can quantify wellness and disease in a more continuous manner, rather than in the current episodic manner. It will allow us to combine data on individual genetics with new human phenotypes at multiple levels, including functional characterization of patient-derived cells, specific physiologic pathways, diet, the microbiome, and wearable physiologic sensors. With its fixed geographic base in New York, the KHP will enable us to measure environmental exposures, including potentially inhaled or ingested toxins, which are currently poorly

[1]*Center for Assessment Technology and Continuous Health, Massachussetts General Hospital, Boston, Massachusetts.*

Some of the material presented in this article has been previously published (Ausiello D, Shaw S. Quantitative human phenotyping: The next frontier in medicine. Trans Am Clin Climatol Assoc. 2014;125:219–226; discussion 226–228) and is used with permission of the author, publisher, and The American Clinical and Climatological Association.

*Address correspondence to: *Dennis Ausiello, Department of Medicine, 55 Fruit Street, Boston, MA 02114,* Email: Ausiello.Dennis@mgh.harvard.edu

measured. This will create another important new data resource that can be integrated with genotypic, phenotypic, and clinical information.

Over time, integrating phenotypes at multiple scales of biology (from cells to the whole individual) will likely become the new norm in human disease studies. This will benefit human health in numerous ways, from individual patient empowerment to biomedical discovery, as well as new approaches to diagnosis, prevention, and precision therapeutics.

In the remainder of the article we first discuss key advances in biological understanding and the failure as yet to realize the full potential of these discoveries to improve human health. We then sketch out the need for new behavioral measurements to advance understanding of the bio-behavioral complex followed by a discussion of how the design of the KHP can liberate such advances in measurement. We conclude by describing how the first implementation of such an endeavor—the newly formed Center for Assessment Technology and Continuous Health (CATCH) at Massachusetts General Hospital (MGH)—illustrates the high potential that increased monitoring has for improving health outcomes.

Human Biology, Genetics, and Health

Advances in human genetics (especially genome-wide association studies) have identified unprecedented numbers of chromosomal loci that contribute to human traits and risk of disease.[1] These genetic data, combined with insights from basic, hypothesis-driven laboratory research, provide a much clearer outline of the genes that contribute to disease (the parts list), including many that were previously unsuspected. The discovery of so many disease loci promises to remake our understanding of disease mechanisms and susceptibility. Particularly exciting is the prospect that new biomarkers of disease risk can be identified.

While the mapping of the genome is the most widely recognized advance in our understanding of human biology, it is but one piece of the human puzzle—there are many other ongoing advances happening, as well. One key scientific advance lies in the emergence of the human microbiota as an important contributor to many chronic diseases. The community of approximately 10^{14} bacterial, archaeal, fungal, and viral cells or particles that reside on each individual constitutes the human microbiota; the "microbiome" additionally refers to the genetic materials and product biomolecules of the microbiota. Microbes colonize the gut at birth and the community is shaped by diet, hygiene, infections, drugs, other environmental exposures, and host genetics. Recent studies in normal volunteers and disease cohorts (such as those with inflammatory bowel disease, obesity, diabetes, or cardiovascular disease) are revealing how all of these genetic and environmental factors can shape what types of microbes are present, and their aggregate influence on human metabolism and immunity (and even behavior in some animal models). (See refs.[2–8] for more study details.)

A third set of advances relates to our understanding of biological pathways. For example, a wide variety of inflammatory cells and pathways are being studied in auto-inflammatory disease as well as common diseases such as type 2 diabetes and cardiovascular disease. Biomarkers such as C-reactive protein or the erythrocyte sedimentation rate are nonspecific markers of inflammation currently used in clinical practice. A variety of new approaches have the potential to enable scientists to parse inflammation more precisely, including serum levels of specific cytokines or mediators, or assays of the activity of inflammatory cells (including molecular imaging of inflammatory cells, or microfluidic devices that can trap or analyze single cells).

In sharp contrast to the ongoing revolution in biological understanding, translation into medical practice has been taking place at a far slower pace. This is particularly true for the clinical translation of genetic discoveries into new approaches to diagnosis, prevention, and treatment. For the vast majority, there has been little to no improvement in treatment or in health. Consider the cases of type 2 diabetes and lipid disorders, each of which is now associated with many dozens of chromosomal regions that influence disease risk. Despite this, these conditions are largely monitored using blood tests that have been used for decades: glycosylated hemoglobin and blood glucose for diabetes, and a lipid panel consisting of fasting total cholesterol, low-density lipoprotein, high-density lipoprotein, and triglycerides.

In another revealing example, inclusion of a genetic risk score (based on several validated variants from genome-wide association studies) failed to improve the ability to predict 10-year risk of developing cardiovascular disease compared to traditional assessments (including factors such as age, gender, smoking history, and the presence of diabetes, hypercholesterolemia, hypertension, or a family history of cardiovascular disease).[9] Despite the revolutions that are taking place in

our biological understanding, the available methods used to diagnose and quantitate disease have conspicuously lagged.

More broadly, while the large number of chromosomal loci newly implicated in many diseases represents a true scientific *tour de force* with tremendous future potential applications in medicine, it remains a challenge to effectively use genetic information to stratify risk. This is likely due to several factors, including the sheer number of genetic loci that can contribute to individual risk and, perhaps most critically, the importance of largely unmeasured environmental and behavioral factors that influence risk of important conditions such as obesity, type 2 diabetes, cardiovascular disease, and cancer. The comprehensive approach to information gathering to be employed by the KHP offers the opportunity to overcome this barrier.

The Bio-Behavioral Complex

A key problem holding back medical advances is that any disease trait is under the influence of not just many genetic loci, but also environmental and behavioral influences. Our understanding of and ability to measure behavior has not advanced at anywhere near the same pace as our ability to understand and measure biological factors. Hence, there is a fundamental need for new approaches to measure human health to better quantify wellness and disease in a more continuous manner as our patients lead their daily lives.

The high potential of the KHP lies precisely in its focus on measuring not just biology but also behavior and the interactions between them. Just as genotyping and genetic sequencing technology have progressed rapidly, an analogous renaissance for phenotypes is required to enable human measurements with greater physiologic resolution and lower cost. For instance, the process of deciphering the physiologic and health consequences of disease-related genetic risk alleles is laborious, expensive, and largely limited to existing phenotypes investigated in small clinical studies. Similarly, novel therapeutic agents are being developed with unprecedented mechanisms but are often evaluated using outdated phenotypes. Novel phenotypes are needed that are more specific and proximate to the mechanisms being modulated. This will enable more rapid testing of novel therapeutic hypotheses in humans, and thus earlier views on the potential efficacy of new agents. Therapeutic trials would also benefit from better stratification of patient subsets to enrich trial populations for those most likely to respond. Stratification by specific genetic mutations has enabled dramatic progress in targeted therapies in certain types of malignancies, such as nonsmall cell lung cancer bearing mutations in the EGF receptor or the ALK kinase. However, for the majority of genetically complex and chronic diseases that are not driven by somatic mutation (as these specific cancer subtypes appear to be), the optimal stratification is likely to come from a combination of genetic and phenotypic stratification.

A new approach to human measurements can also transform how individuals engage in their own health, provide insightful measurements in real time, and allow individuals to monitor and improve their own health and wellness (in partnership with their physicians and caregivers). A key challenge is to move the monitoring of health and disease away from the physical and time constraints of physician offices and hospitals, and into the domain of patients' lives. The ability to track symptoms or health status more quantitatively can help patients and their caregivers understand disease trends and how interventions may worsen or ameliorate symptoms, and allow the time during an office visit to be used more effectively.

Another group of poorly measured factors relates to diet and environmental exposures, which may potentially contribute inhaled or ingested toxins. Exposures are typically accessed on rare occasions through survey instruments based on recall, blood, or urine assays. While continuous measurement of environmental exposures may not be necessary (or feasible), enabling more facile and passive quantification of environmental exposures will create an important new data resource that can be integrated with genetic and clinical information.

Continuous Measurement, Big Data, and the KHP

Our current system of delivering healthcare is episodic and reactive. That is, patients see their physicians largely at regularly scheduled intervals (typically one year), and/or when symptoms appear or worsen. At a time when the healthcare system in the United States faces tremendous pressure to contain costs and improve efficiency and outcomes, this episodic approach forces patients to summarize and communicate months of symptoms and observations in a brief office visit, and limits the ability of patients and physicians to proactively address emerging medical issues.

During their episodic appointments, the methods physicians use to assess disease in our patients have largely remained the same for decades. The typical

office visit will document the patient's medical history and symptoms since the last visit (usually several months ago); parameters such as weight, heart rate, blood pressure, and respiratory rate; a physical examination; and perhaps standard blood tests such as general chemistry values and a lipid panel. While specialized blood diagnostics and imaging studies are used to investigate specific diagnoses, the most commonly used measures reflect an uneasy balance between cost, the time constraints of an office visit, and the ability to detect significant changes in health status.

We are fortunate to live in an era in which increased behavioral and biological measurement is technologically possible. The increasing availability of digital and genetic data, and measurement platforms thereof, facilitates the collection and analysis of extremely large datasets containing multiple types of data relevant to human health. These include the data contained in electronic medical records (EMRs), genetic data (e.g., characterization of mutations in a tumor biopsy, or genome-wide genotyping), and pharmacy or claims data. Less traditional but growing sources of data include personal fitness trackers (such as wearable activity monitors), online social communities and other web forums, and medical measurements such as blood pressure or blood glucose that can be transmitted wirelessly. Growing analytic and computing capabilities are enabling mining of these large datasets for insights at the level of individual patients or entire populations, and can have a profound impact on how healthcare is administered in the United States in the future.

Certain types of medical data are already collected continuously, and represent opportunities for data repurposing. For instance, millions of implanted cardiac devices such as pacemakers and implantable cardioverter-defibrillators provide ready access to beat-by-beat heart rate data. Continuous glucose monitors (typically accessed via a small sensor in the interstitial space) have long been used primarily to guide dosing of automated insulin infusion pumps, but may yield insights into the dynamics of glucose regulation.

Other important behaviors to measure include exercise, diet, and medication adherence and make significant contributions to several diseases ranging from cancer to diabetes and cardiovascular disease. Data from wearable devices such as activity monitors (such as digital pedometers) and wrist-based monitors (e.g., devices that measure skin galvanic response as a reflection of stress) could provide insight into individual behaviors as well as facilitate feedback. Several types of wearable measurements represent physiologic parameters that could be analyzed in certain disease-specific contexts, but also represent important sources of information for individuals as they monitor their own health.

Studies suggest that individuals commonly discontinue wearable devices after several months, potentially limiting their widespread application. But embedding sensors into devices that are ubiquitous and used with high persistence, such as mobile phones, may open up new avenues. Furthermore, thanks to increased capabilities, mobile phones are increasingly used for a variety of routine behaviors such as communication, travel location, and even specific health-related software ("apps"), and the mobile phone represents an appealing platform for a variety of continuous and behavioral measurements.

But the biggest opportunity lies in the fact that we are currently in the midst of unprecedented technological change. A profusion of powerful computing and communications platforms has been enabled by small form factors and ubiquitous Internet access. These include smart phones or tablet computers with cellular network and/or wireless Internet access, and personal devices that can transmit digital health-related information. These devices and associated software or apps place unprecedented data collection, retrieval, and exchange capabilities literally in the hands of individuals, and liberate these activities from traditional location-based constraints (such as physician offices or hospitals). However, these powerful capabilities are only beginning to be systematically explored in the context of individual health.

The importance of the ongoing mobile revolution is that it makes the gathering of measurements more unobtrusive to the individual. Intermediate benefits may be realized by current "mobile health" efforts that use traditional measurements such as blood glucose or weight and simply transmit them to caregivers (e.g., through an iPhone attachment that measures blood glucose or wireless-equipped weight scales). But a full realization will require quantitative measurements that can be collected passively. This has spawned great interest in the so-called wearable sensors, such as devices worn on the waist or wrist, embedded in smart phones, or even embedded in clothing that could reflect physiologic parameters such as heart rate and respiration, activity, stress, or behavior. Passive data collection significantly increases the completeness of data captured about the populations to which this approach may be applied; active data entry risks biasing the study population toward

participants who have higher levels of technological familiarity or higher degrees of motivation or engagement in their health. More complete, less biased datasets will better allow analysis to reveal the effect of therapeutic or other interventions.

The KHP will liberate progress by convening a wider range of expertise than is traditional, including device engineers, front-line physicians, geneticists, and experts in behavioral modeling focused directly on the development of novel phenotypes. By tracking individuals in rich detail over a long period of time, the KHP will capture and catalog dynamic patterns of individual and social behavior in new and richer metrics than ever before. This will help us develop new approaches to measuring human health, and so we can quantify wellness and disease in a more continuous manner, rather than in the current episodic manner.

With its fixed geographic base in New York, the KHP will enable us to measure environmental exposures, including potentially inhaled or ingested toxins, which are currently poorly measured, providing consistency across the study population that could not be achieved with a more disaggregated geographic study frame. This will create another important new data resource that can be integrated with genotypic, phenotypic, and clinical information.

Given the multiple modalities of measurement that it is designed to include, the KHP will liberate studies that combine data on individual genetics with new human phenotypes at multiple levels, including functional characterization of patient-derived cells, specific physiologic pathways, diet, the microbiome, and wearable physiologic sensors. Because of the diversity of this universe of potential measurements, the development of integrative analytics that allow disparate data types (traditional and nontraditional) to be analyzed in concert will be critical. Integrating phenotypes at multiple scales of biology (from cells to the whole individual) will likely become the new norm in human disease studies.

Continuous Monitoring, CATCH, and the Future of Healthcare

An ongoing effort taking place at MGH illustrates the high potential that increased monitoring has to change the current episodic healthcare paradigm and improve health outcomes. The newly formed CATCH seeks to discover and apply new ways to quantitatively measure human phenotypes in health and disease. Through a multidisciplinary collaboration of scientists, physicians, engineers, computer scientists, and behavioral experts across MGH, the Massachusetts Institute of Technology, and the private sector, CATCH will leverage the digital and genetic revolutions to transform how individuals monitor their own health, and how physicians can prevent, diagnose, and treat disease.

To fully implement this vision, important changes are also needed in the culture of patient care and scientific research. Scientific collaborations will need to convene a wider range of expertise than is traditionally sought, including device engineers, front-line physicians, geneticists, and experts in sociology and behavior. Collection of these novel data types will require new approaches to data ownership and security that appropriately balance an individual's control over use of their data and with permission and trust framework for secondary use of data in specific contexts. Analyses must be focused on actionable insights and rendered visually to allow patients and their caregivers to understand the medical implications.

CATCH provides an important model for the collection of comprehensive phenotypic data, demonstrating the promise of this approach. The KHP will capitalize and expand upon the ideas and lessons coming out of CATCH in order to successfully characterize the bio-behavioral complex on a large scale. Together, these two projects will enhance our understanding of the complex forces that shape human health.

Implementation in the KHP

Investigations that are outlined in this article would utilize the following KHP data sets, among others:

1. Medical information on study participants' health would be available from the medical history and records going forward (EMRs, doctor's notes, hospital records, dental records). Prescription data would be gathered via the NY State Prescription database. This information would be complemented by the SPARCS database and KHP's own tests: blood tests (blood metals, vitamins, lipids, glucose, and other biomarkers), and urine and hair tests (smoking, alcohol, and substance use) every three years.
2. Information on genetics would be gathered via whole genome sequencing of blood samples for adults (saliva for children) performed at study intake. In addition, data on epigenetics would be gathered via triennially performed assays.
3. Microbiome (oral and gut) data would be gathered via periodic saliva and stool samples. In addition,

these samples would be targeted for collection during each health or emotional crisis.
4. Dietary data would be collected via periodic food diaries, complemented by mining for food purchases in financial data (credit cards, debit cards, checks).
5. Mobility and activity level information would be collected by activity trackers and smartphone apps.
6. Information on financial status and participation in government assistance programs (SNAP, Social Security, TENF) would be available via financial data gathered using a combination of automated and survey-based methodologies.
7. Connectivity to existing and future medical devices would be gauged through the use of technologies that are most widely used for device-to-device connectivity. Initially, KHP smartphone app would connect to any Bluetooth-enabled device to collect device data.

Conclusions

Ultimately, traditional clinical information must be combined with genetic data and nontraditional phenotypes, and analyzed in a manner that yields actionable insights into disease diagnosis, prevention, or treatment. Real-time, quantitative human phenotyping and associated analytics will enable individuals, caregivers, and scientists to better quantify wellness and disease in a more continuous manner, and as individuals lead their daily lives. The bio-behavioral complex is the next great biomedical frontier, analogous to the Human Genome Project in its profound implications for medicine, and in the scale of the effort and resources required. With the knowledge generated by CATCH and the KHP, we will finally close the gap between biological advances and healthcare practice.

Author Disclosure Statement

No competing financial interests exist.

References

1. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–D1006.
2. Karlsson F, Tremaroli V, Nielsen J, Backhed F. Assessing the human gut microbiota in metabolic diseases. Diabetes. 2013;62:3341–3349.
3. Karlsson FH, Fak F, Nookaew I, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat Commun. 2012;3:1245.
4. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: Human gut microbes associated with obesity. Nature. 2006;444:1022–1023.
5. Karlsson FH, Tremaroli V, Nookaew I, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498:99–103.
6. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490:55–60.
7. Kostic AD, Gevers D, Siljander H, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015;17:260–273.
8. Cho I, Blaser MJ. The human microbiome: At the interface of health and disease. Nat Rev Genet. 2012;13:260–270.
9. Thanassoulis G, Peloso GM, Pencina MJ, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: The Framingham Heart Study. Circ Cardiovasc Genet 2012;5:113–121.

Cite this article as: Ausiello D, Lipnick S (2015) Real-time assessment of wellness and disease in daily life. *Big Data* 3:3, 203–208, DOI: 10.1089/big.2015.0016.

Abbreviations Used

CATCH = Center for Assessment Technology and Continuous Health
EMRs = electronic medical records
MGH = Massachusetts General Hospital

Big Data
Volume 3 Number 3, 2015
DOI: 10.1089/big.2015.0015

ORIGINAL ARTICLE

Opportunities for New Insights on the Life-Course Risks and Outcomes of Cognitive Decline in the Kavli HUMAN Project

Kenneth M. Langa[1,*] and David Cutler[2]

Abstract

The Kavli HUMAN Project (KHP) will provide groundbreaking insights into how biological, medical, and social factors interact and impact the risks for cognitive decline from birth through older age. It will richly measure the effect of cognitive decline on the ability to perform key activities of daily living. In addition, due to its family focus, the KHP will measure the impact on family members, including the amount of time that family members spend providing care to older adults with dementia. It will also clarify the division of caregiving duties among family members and the effects on caregivers' work, family life, and balance thereof. At the same time, for care that the family cannot provide, it will clarify the extent to which cognitive decline impacts healthcare utilization and end-of-life decision making.

Key words: aging; cognitive decline; population research

Introduction

Dementia is a common and feared aging-related condition characterized by declines in memory and other cognitive functions that are severe enough to cause the loss of independent function and difficulties with activities of daily living. Dementia has a large and growing impact on older adults, their families, and government programs in the United States and around the world. About 4.2 million adults in the United States, and more than 35 million worldwide, had dementia in 2010,[1] with an estimated economic impact in the United States of about $200 billion[2] and $600 billion worldwide,[3] including a large burden of unpaid caregiving provided by families. Because of the sharp increase in the incidence of dementia at older ages and the expected growth in the worldwide elderly population in the decades ahead (from about 600 million in 2015 to 1.5 billion in 2050), the number of dementia cases is expected to triple by 2050, absent new interventions to prevent or slow the trajectory of cognitive decline.[1,4]

Recognition of the growing impact of dementia has led governments around the world to prioritize expanding the collection of data on individuals and populations to understand, address, and track better the current and future impact of the dementia epidemic. For instance, the National Alzheimer's Project Act (NAPA) was signed into law by President Obama in 2011 in order to expand U.S. government efforts to improve treatment and prevention, and to collect data to track progress of these efforts in the future. The G8 Dementia Summit was held in London in 2013 in recognition of the growing global impact of Alzheimer's disease (AD) and dementia, and to begin to coordinate efforts for international collaboration and data sharing. The World Health Organization also recently identified dementia as a "public health priority" that should be on all countries' public health agenda.[4]

While the large growth in the number of older adults in the coming decades will lead to an increase around the world in dementia cases, a number of recent studies have suggested that the age-specific risk of dementia has actually decreased in high-income countries over the last 25 years, possibly due to more aggressive

[1]*University of Michigan, Ann Arbor, Michigan.*
[2]*Harvard University, Cambridge, Massachusetts.*

*Address correspondence to: *Kenneth M. Langa, University of Michigan, 2800 Plymouth Road, Bldg 16, Room 430W, Ann Arbor, MI, 48109,* E-mail: klanga@umich.edu

treatment of cardiovascular risk factors that increase the risk of cognitive decline (e.g., diabetes, hypertension, and high cholesterol), as well as a worldwide boom in educational attainment, which is thought to increase the "cognitive reserve" of older adults.[5] However, it is unclear whether the recent large increase in the prevalence of obesity and diabetes around the world will cause a slowing or reversal in this optimistic trend, and it is also unclear whether there has been a similar or opposite trend in low- and middle-income countries.[6,7]

Over the last two decades, the conceptualization of the cause for cognitive decline and dementia in older adults has evolved from a focused derangement of a few biological pathways (e.g., the amyloid cascade of AD; multiple strokes causing vascular dementia) toward a fuller life course, multifactorial syndrome that results in overlapping brain pathologies ("mixed dementia" from amyloid plaques, vascular insults, and Lewy-body pathology) due to both biological and social stressors from the womb through old age. In this way, dementia is a prototypical health condition that requires a "life-course" perspective because of the exquisite sensitivity of the development of brain cells and the healthy wiring of brain networks to both daily biological and social stressors.[8]

The intensive longitudinal daily data collection planned for the Kavli HUMAN Project (KHP) will provide new opportunities for ground-breaking insights into how biological, medical, and social factors interact to increase or decrease the risks for cognitive decline from birth through older age, including: (a) the transmission of "brain health" across generations (e.g., maternal health and brain gestation; early parent–child social interactions and the building of "cognitive reserve"); (b) how education, cognitive stimulation, and daily cognitive activity (reading text, smartphones, the Internet, etc.) are linked with brain health and the risk for later-life cognitive decline; (c) how key cardiovascular risk factors, such as hypertension, diabetes, vascular disease, and physical inactivity, affect the risk for cognitive decline; and (d) how social life and social interactions are related to brain health and the risk for cognitive decline.

In addition to providing new and unique data regarding the complex interaction among biological, medical, and social risk factors for cognitive decline, the KHP will provide an unprecedented opportunity to understand better the wide-ranging impact and outcomes of cognitive decline on individuals, their families, and the wider social networks in which they live their daily lives, including: (a) changes in the ability to perform key activities of daily living; (b) the amount of time that family members spend providing daily care to older adults with dementia, as well as the trajectory caregiving time as cognitive impairment worsens; (c) the dynamics of the division of caregiving duties among family members, and how caregiving affects the work and family life of caregivers; and (d) healthcare utilization and decision making at the end-of-life.

There are many key measurement opportunities and challenges involved in developing the ideal data set. A key set of factors to be measured are maternal activities during gestation. There are also important early-life interactions and opportunities to be measured. For example, more educated parents likely speak with their children in different ways than less educated parents, perhaps leading to different levels of "cognitive reserve" and, therefore, different levels of risk for late-life cognitive decline. In regard to measurement, a number of studies have used in-home recordings of parent–child conversations to identify the quantity and "quality" of cognitive stimulation for kids. It is also increasingly possible to track daily cognitive activity (reading text, smartphone use, Internet use, etc.), including how labor force participation affects risk for cognitive decline in later life.

Another important set of risk factors to measure relate to physical activity and biological markers. In addition to simple measures of how sedentary individuals are, it is important to track exercise patterns. Going further, it is important to measure such cardiovascular risk factors by monitoring blood pressure, related diseases, and the extent to which programs to control these are implemented (diabetes control, etc.). Other risk factors to measure include dietary choices, alcohol use, and use of other substances that are psychoactive and may contribute either positively or negatively to cognitive abilities in later life. It is also important to track mental health directly, including ongoing anxiety, depression, and more severe conditions, along with the measures and substances taken in an effort to rectify such conditions.

Given that the ultimate object of study is cognitive decline, it is also important to monitor and improve understanding of possible early warning signals. These may include changes in financial activities and patterns of decision making, such as "giving up the checkbook." There may be a concomitant pattern of increased risk of being a victim of financial fraud, so that key financial transactions must be monitored. There

are also possible telltale patterns of movement, such as constriction of daily geography (e.g., not venturing far from familiar places). It is also important to track changes in the frequency and types of social interactions, as well as changes in speech patterns.

Implementation in the KHP

A key advantage of the holistic approach that KHP liberates is the ability to track the impact of cognitive decline on a large number of outcome variables. In the medical arena, it is important to track how cognitive decline is caused by and contributes to general health, function, and well-being in later life. It is also important to monitor the impact of cognitive decline on families and social networks. Of particular interest is the impact on labor-force participation of individuals and their caregivers. It is also important to track the economic cost of healthcare utilization, long-term care, and informal caregiving.

To achieve these ambitious targets, the KHP will develop a rich set of methods for measuring and tracking cognitive function. It is proposed that standard cognitive screening batteries be undertaken at yearly intervals. This will be combined with daily measurement of aspects of cognitive function with smartphone apps and other monitoring devices. GPS information will be used to gauge how daily "life space" changes as cognition declines. There may also be occasion to conduct explicit brain images at regular intervals and at key moments of change. Finally, there may be need for proxy (informant) assessments of cognitive and other functions as individuals become less able to participate in survey or other self-report data collection.

Overall, investigation of factors that contribute to cognitive decline from birth through older age would utilize the following KHP data sets, amongst others. (a) Medical information on the mother's health, pregnancy, and birth of the baby would be available from the medical history and records going forward (EMRs, doctors' notes, hospital records, dental records). Prescription data would be gathered via the New York State Prescription database. This information would be complemented by the SPARCS database and the KHP's own tests: blood tests (blood metals, vitamins, lipids, glucose, and other biomarkers) together with urine and hair tests (smoking, alcohol, and substance use) every 3 years. (b) Exposure to toxins and other chemicals would be measured via silicone wristbands worn periodically. (c) Formal education data would be available via participants' record cards, standardized test results (e.g., No Child Left Behind, ACT, SAT), and New York City Department of Education databases on student progression and school rankings. (d) Information on informal education would be gathered by KHP field teams via questionnaires on extracurricular activities and by counting the number of books in participants' homes. (e) Time parents spend together with children would be measured using Bluetooth-based presence sensors in the home and by smartphone apps that measure social interactions. (f) Data on caregiving by family members would be available via questionnaires, Bluetooth-based presence sensors in the home, and a smartphone app that measures social interactions. (g) Mobility and activity level information would be collected by activity trackers and smartphone apps. (h) Time spent on digital devices would be measured by apps on smartphones, tablets, and PCs. (i) Dietary data would be collected via periodic food diaries, complemented by mining for food purchases in financial data (credit cards, debit cards, checks). (j) Information on financial status and participation in government assistance programs (SNAP, Social Security, TENF) would be available via financial data gathered using a combination of automated and survey-based methodologies. (k) Information on smoking and alcohol use would be collected on an ongoing basis by mining for purchases of tobacco products and alcoholic beverages in financial data, in addition to the triennial tests using biological samples.

The impact of various factors on cognition and genetics would be analyzed via the following KHP data sets, amongst others. (a) Cognitive function levels would be measured via self-administered psychology questionnaires/tests on smartphones and tablets at intake and periodically thereafter. (b) Information on genetic variation—individual and family—would be gathered via whole genome sequencing of blood samples for adults (saliva for children) performed at study intake. In addition, data on epigenetics would be gathered via triennially performed assays. (c) "Giving up on the checkbook" could be measured indirectly via analyzing trends on the quantity and types of financial transaction individuals perform over time.

Author Disclosure Statement

No competing financial interests exist.

References

1. Prince M, Guerchet M, Prina M. Policy Brief for Heads of Government: The Global Impact of Dementia 2013–2050. London: Alzheimer's Disease International, 2013.

2. Hurd M, Martorell F, Delevande A, et al. The monetary costs of dementia in the United States. N Engl J Med. 2013;368:1326–1334.
3. Wimo A, Jönsson L, Bond J, et al. The worldwide economic impact of dementia 2010. Alzheimers Dement. 2013;9:1–11.
4. Langa KM. Is the risk of Alzheimer's disease and dementia declining? Alzheimers Res Ther 2015;7:34.
5. Larson EB, Yaffe K, Langa KM. New insights into the dementia epidemic. N Engl J Med. 2013;369:2275–2277.
6. Chan KY, Wang W, Wu JJ, et al. Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990–2010: a systematic review and analysis. Lancet 2013;381:2016–2023.
7. Wu YT, Lee HY, Norton S, et al. Prevalence studies of dementia in mainland china, Hong Kong and Taiwan: a systematic review and meta-analysis. PLoS One 2013;8:e66252.
8. Lindenberger U. Human cognitive aging: corriger la fortune? Science 2014;346:572–578.

Cite this article as: Langa KM, Cutler D (2015) Opportunities for new insights on the life-course risks and outcomes of cognitive decline in the Kavli HUMAN Project. *Big Data* 3:3, 189–192, DOI: 10.1089/big.2015.0015.

Abbreviations Used

NAPA = Nahcnal Alzheimer's Project Act
AD = Alzheimer's disease
KHP = Kavli Human Project

Big Data
Volume 3 Number 3, 2015
DOI: 10.1089/big.2015.0013

ORIGINAL ARTICLE

How Genetic and Other Biological Factors Interact with Smoking Decisions

Laura Bierut,[1,*] and David Cesarini[2]

Abstract

Despite clear links between genes and smoking, effective public policy requires far richer measurement of the feedback between biological, behavioral, and environmental factors. The Kavli HUMAN Project (KHP) plans to exploit the plummeting costs of data gathering and to make creative use of new technologies to construct a longitudinal panel data set that would compare favorably to existing longitudinal surveys, both in terms of the richness of the behavioral measures and the cost-effectiveness of the data collection. By developing a more comprehensive approach to characterizing behavior than traditional methods, KHP will allow researchers to paint a much richer picture of an individual's life-cycle trajectory of smoking, alcohol, and drug use, and interactions with other choices and environmental factors. The longitudinal nature of KHP will be particularly valuable in light of the increasing evidence for how smoking behavior affects physiology and health. The KHP could have a transformative impact on the understanding of the biology of addictive behaviors such as smoking, and of a rich range of prevention and amelioration policies.

Key words: smoking; genetics; deep phenotyping; smoking cessation

Introduction

Despite the fact that it has long been understood that smoking is a leading modifiable risk factor for poor health,[1] estimates suggest that tobacco use continues to be responsible for nearly one in five U.S. deaths.[2] Even though the development of smoking cessation and prevention strategies has been a major priority for policy makers for quite some time, progress has been hampered by our as-of-yet imperfect understanding of the complex genetic and environmental etiology of smoking behavior. In an era of rapid technological advances in the measurement and analysis of DNA, the understanding of robustly established—but difficult-to-interpret—genetic associations with smoking behavior made possible by "big data" can be substantially enhanced through careful follow-up analyses in rich longitudinal panels with data of high quality.

The advent of genome-wide association studies (GWAS) has massively increased the ability to identify genes that impact deleterious behaviors. In particular, robust and biologically plausible associations have been discovered between smoking and genes. Yet, different facets of smoking behavior—initiation, intensity, and cessation—have distinct biologic and environmental contributors. To test hypotheses about genetic effects on smoking, it is therefore critical to have reliable measures of the various facets of smoking behavior over the life cycle. Yet, current behavioral measures, such as maximum level of smoking at any point in the life cycle, remain crude.

By radically improving measurement of behavioral phenotypes, the Kavli HUMAN Project (KHP) will clarify links between biology and health-impacting behaviors such as smoking. For example, self-reported smoking quantities can be cross-checked against credit-card records on cigarette purchases and supplemented with information from medical records about health conditions associated with tobacco use. Direct biological measurement of smoking markers, such as cotinine—a compound formed after nicotine enters the body—and exhaled carbon monoxide—a measure

[1]*Washington University in St. Louis, St. Louis, Missouri.*
[2]*New York University, New York, New York.*

*Address correspondence to: *Laura Bierut, Washington University in St. Louis, 660 South Euclid, St. Louis, MO 63110*, E-mail: bierutl@psychiatry.wustl.edu

of exposure to smoked combustible cigarettes—will also be informative.

The KHP will particularly enrich the understanding of feedback mechanisms between biology and behavior. For example, studies have begun to identify several genes whose levels of methylation are associated with smoking behavior. Whether these changes can help explain some of the biological pathways through which smoking ultimately impacts lung health and lung cancer is a vibrant area of research. Longitudinal data sets with rich behavioral and biological measures can be an invaluable resource for enhancing the understanding of the links between smoking and health.

Genetics of Smoking

Beginning around 2005, medical genetics research began to undergo a paradigm shift, moving to GWAS. In these studies, made feasible by technological advances, researchers test the outcome of interest for association with each of the measured single-nucleotide polymorphisms (SNPs). Because of the large number of hypotheses tested in a GWAS, a SNP association is considered to be established only if it reaches the "genome-wide significance" threshold of $p < 0.00000005$. Adequate statistical power at this stringent significance threshold requires very large samples. Since individual samples are generally too small, many GWAS are conducted within research consortia that meta-analyze results from multiple samples and countries.

Empirically, it is now well established that results from such GWAS replicate very consistently.[3] There are several reasons for the robustness of GWAS findings (see Rietveld et al.[4] for a discussion). Before the modern era of GWAS, most molecular genetic studies of smoking had been candidate gene studies, which focused exclusively on studied variations in genes in biological systems known to play an important role in nicotine addiction. The replication record of these early studies turned out to be disappointing, and the estimates of the effect sizes were often highly heterogeneous across studies.[5] An influential review[6] concluded that the "evidence for a contribution of specific genes to smoking behavior remains modest." Ten years later, the GWAS have uncovered a handful of genetic associations with smoking behavior for which the evidence is very strong and the replication record is excellent.

A landmark event in the study of the genetics of smoking was the publication of the first GWAS of smoking in *Nature*,[7] along with two studies of lung cancer in *Nature*[8] and *Nature Genetics*.[9] This work was followed by three GWAS of smoking behavior in the May 2010 issue of *Nature Genetics*.[10–12] By far the strongest results came from a set of SNPs located in the chromosome 15 cluster of virtually adjacent nicotinic receptor genes (*CHRNA3*, *CHRNA5*, and *CHRNB4*), which were identified in all studies as a risk factor for heaviness of smoking defined by number of cigarettes smoked per day (CPD), as well as the strongest genetic risk for the development of lung cancer. The SNP rs16969968, known colloquially among researchers as "Mr. Big," is widely believed to be the causal variant underlying the signal. In particular, it is known to cause an amino acid change in the alpha-5 subunit of the nicotinic receptors, and experiments have found that this change alters the responsiveness of the nicotinic receptors to nicotine.[13]

Despite many strengths, the GWAS also have some obvious limitations. First, it is often necessary to sacrifice phenotype quality to attain sample sizes needed for studies to have adequate power to detect associations. As a result, it is not always easy to interpret an observed association. For example, the "TAG" study[11] combined quite different measures in a single study: some cohorts asked smokers about their maximum daily consumption at peak consumption, whereas others asked about contemporaneous consumption. Moreover, GWAS are useful for identifying genetic signals, but are of limited value for understanding how an identified genetic effect might vary across environmental conditions. In the case of smoking, it is *a priori* plausible that such interactions are often of first-order importance.

Thus, GWAS are of great value for detecting real and replicable genetic associations, but they are merely a necessary first step toward the more ambitious twin goals of identifying the ensemble of genes that, along with environmental factors, account for heterogeneity across individuals, and understanding how environmental factors can amplify or dampen genetic risk. Credibly establishing such pathways requires rich longitudinal measures of behavior, biological markers, and environmental factors. Because no such data set presently exists, the KHP could potentially fill an important void.

Figure 1 shows why it is likely that this void will continue to grow in the coming years, as larger and larger discovery samples lead to the discovery of more and more genetic associations with various complex outcomes. For example, the first study of schizophrenia identified a single polymorphism,[14] but the availability of larger and larger samples has brought the number up to 108.[15] Early studies of height identified 10–20

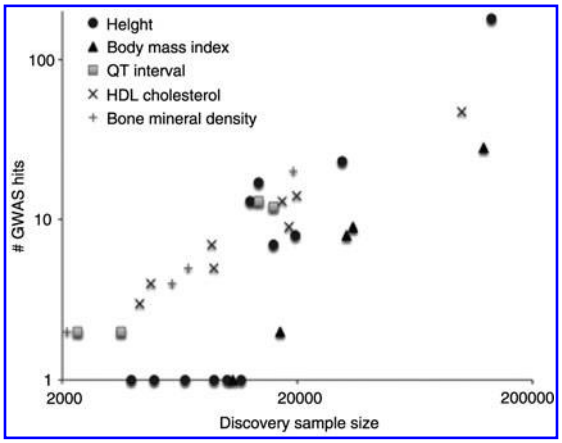


FIG. 1. Number of polymorphisms identified as a function of meta-genome-wide association studies sample size. Note the logarithmic scale on the *x*-axis. (Source: Visscher et al.[3]).

polymorphisms,[16–18] whereas new research by the GIANT consortium,[17] based on a sample of 250,000 individuals, identifies 700.

It seems exceedingly likely that in the coming years, we will similarly be awash in genetic associations with smoking phenotypes as well as measures of other substance use. To have the greatest scientific impact, these associations will require interpretation and follow-up work on behavioral and biological mechanisms. Below, three concrete examples are given of the complementarities believed to exist between the use of "big data" for gene discovery and the use of high-quality longitudinal data with rich behavioral, biological, and environmental measure to refine the understanding of mechanisms.

Improving Phenotype Measurement by Leveraging Multiple Data Sources

A major challenge in addiction research is phenotype measurement. By combining conventional longitudinal survey-based measures with novel ways of measuring smoking behavior, the KHP will allow researchers to paint a much richer picture of an individual's life-cycle trajectory of smoking, alcohol, and other substance use. For example, measuring substance use behavior is associated with three major difficulties. First, subjects who are surveyed on a single occasion may exhibit recall biases. Second, because of the social stigma associated with substance use, some respondents may systematically color their responses. Third, many surveys ask about substance use at a single point in time, and responses to these questions may be poor proxies for an individual's life cycle of substance use behavior.

The KHP data could be leveraged in a number of ways to obtain more reliable measures of substance use. By tracking people longitudinally, it is possible to measure changes in substance use behavior much more accurately over time. Moreover, a number of other data sources could be used to improve phenotype measurement and validate survey responses. For example, self-reported smoking quantities could be cross-checked against credit-card records on cigarette purchases. There are also a number of well-known biomarkers for smoking behaviors and other substance use. Cotinine, which can be measured from saliva,[19] is often used to obtain an objective measure of an individual's exposure to tobacco.[20] An exciting development in recent years is the fact that it is becoming feasible to measure DNA methylation, an epigenetic mechanism for the regulation of gene expression. Epigenome-wide association studies have identified several genes whose methylation is strongly associated with smoking behavior.[21] Finally, survey questions could be supplemented with information from medical records about health conditions associated with tobacco use (such as diagnostic codes for pulmonary disease and lung cancer) or diagnostic codes for treatment of tobacco use and dependence.

Existing genetic studies suggest that the genetic architecture of different facets of smoking behavior—initiation, intensity, cessation—show quite modest genetic overlap. To test hypotheses about genetic effects on smoking, it is therefore critical to have reliable measures of the various facets of smoking behavior over the life cycle.

Illuminating Biological Consequences of Health Behaviors

A very robust finding emerging from the epigenome-wide association studies of methylation conducted to date is that smoking is associated with the methylation of many genes. Whether these methylation patterns can help to explain some of the biological pathways through which smoking ultimately impacts lung health[22] and lung cancer[23] is a vibrant area of research. The KHP would be a valuable resource for testing hypotheses about several of the genes whose methylation is believed to play an important role in the causal pathways from smoking to poor health. Most studies measure methylation from the blood, but methylation can also be measured in other types of tissue, including saliva, which is easier and cheaper to collect on a large scale.

Gene–Environment Interactions and Behavioral Pathways

Finally, the KHP data could be a valuable resource for testing hypotheses about gene–environment (G×E) interactions. Efforts to understand interactions between environmental factors and tobacco and alcohol consumption are already well underway.[24,25] A major challenge for studies of G×E is that the measures of environmental exposures are often imperfect; the KHP's ambitious plans for gathering high-quality data on life events and other environmental variables would thus fill an important void. A second challenge is that to deliver convincing answers, G×E studies need to have adequate statistical power.[26] The large and richly phenotyped KHP sample would thus help to overcome two serious obstacles to scientific progress in this area. Indeed, the large sample would permit meaningful analyses even in fairly narrowly defined subgroups. Hypotheses about interactions could be tested in suitably selected subsamples through randomized interventions.

In studies of G×E, it is also envisioned that there will be large gains from collaborations between geneticists, who contribute critical biological expertise, and economists, who are well trained in teasing out causal relationship from observational data. In the social sciences, controlled experiments are not always a feasible research strategy for establishing causality. Confronted with this reality, researchers have shown great ingenuity in developing methods to tease out causal relationships from "quasi-experiments," events that produce variation that plausibly resembles the experimental variation generated by a controlled experiment (for an overview, see Angrist and Pischke[27]). For example, studies have studied lottery winners to study the causal impact of wealth on labor supply,[28] and adoptees assigned to families using plausibly random mechanisms to learn about the impact of family environment on child outcomes.[29] During the course of the study, it is likely that some subjects will be exposed to plausibly exogenous environmental insults, for example a large unanticipated bequest or serious injury from an accident. Such naturally occurring variation can be leveraged to gain insight into causal interactions between genetic and environmental factors.

Implementation in the KHP

Investigation of factors that contribute to smoking decisions would utilize the following KHP data sets, among others: (a) Smoking use data would be available via medical history and records forward, and through KHP's biological samples (see below), as well as by mining for purchase of tobacco products in the financial data. Mining financial data would offer additional benefits over the limitations of survey-only methods due to its continuous basis, and it would also help confirm actual cessation of smoking versus "claimed" cessation. (b) Air quality and ambient noise levels would be measured via sensors placed in the home. (c) Exposure to toxins and other chemicals would be measured via silicone wristbands worn periodically. (d) Information on financial status and participation in government assistance programs (Supplemental Nutrition Assistance Program, Social Security, Temporary Cash Assistance to Needy Families) would be available via financial data gathered using a combination of automated and survey-based methodologies.

The impact of smoking decisions on health would be analyzed via the following KHP data sets, amongst others: (a) Medical information on study participants' health would be available from the medical history and records going forward (medical records, doctors' notes, hospital records, dental records). Prescription data would be gathered via the NY State Prescription database. This information would be complemented by the Statewide Planning and Research Cooperative System database and KHP's own tests: blood tests (blood metals, vitamins, lipids, glucose and other biomarkers), urine and hair tests (smoking, alcohol and substance use) every 3 years. (b) Information of a genetic nature, including telomere length, would be gathered via whole genome sequencing of blood samples for adults (saliva for children) performed at study intake. In addition, data on variation in epigenetics would be gathered via triennially performed assays.

Conclusion

It has been emphasized that there is no conflict between research approaches that leverage enormous data sets to discover basic patterns of association and research approaches leveraging rich longitudinal data sets to test specific causal hypotheses. Rather, two approaches should be viewed as mutually reinforcing and necessary for making progress on designing effective health interventions.

Author Disclosure Statement

Laura J. Bierut is listed as an inventor on Issued U.S. Patent 8,080,371, "Markers for Addiction" covering the use of certain SNPs in determining the diagnosis,

prognosis, and treatment of addiction. For David Cesarini, no competing financial interests exist.

References

1. Centers for Disease Control and Prevention. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta, Georgia 2010. Available online at www.cdc.gov/tobacco/data_statistics/sgr/2010/index.htm?s_cid=cs_1843 (last accessed June 1, 2015).
2. Mokdad AH, Marks JS, Stroup DF, et al. Actual causes of death in the United States, 2000. JAMA. 2004;291:1238–1245.
3. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
4. Rietveld CA, Conley D, Eriksson N, et al. Replicability and robustness of GWAS for behavioral traits. Psychol Sci. 2014;25:1975–1986.
5. Aljasir B, Ioannidis JP, Yurkiewich A, et al. Assessment of systematic effects of methodological characteristics on candidate genetic associations. Hum Genet. 2013;132:167–178.
6. Munafò M, Clar T, Johnstone E, et al. The genetic basis for smoking behavior: a systematic review and meta-analysis. Nicotine Tob Res. 2004;6:583–597.
7. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638–642.
8. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008;452:633–637.
9. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. 2008;40:616–622.
10. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42:436–440.
11. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010;42:441–447.
12. Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet. 2010;42:448–453.
13. Bierut LJ, Stitzel JA, Wang JC, et al. Variants in nicotinic receptors and risk for nicotine dependence. Am J Psychiatry 2008;165:1163–1171.
14. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752.
15. Ripke S, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature 2014;511:421–427.
16. Gudbjartsson DF, Walters GB, Thorleifsson G, et al. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008;40:609–615.
17. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 2010;467: 832–838.
18. Weedon MN, Lango H, Lindgren CM, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008;40:575–583.
19. Etter JF, Vu Duc T, Perneger TV. Saliva cotinine levels in smokers and nonsmokers. Am J Epidemiol. 2000;151:251–258.
20. Benowitz NL. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev. 1996;18:188–204.
21. Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. Front Genet 2013;4:132.
22. Zong DD, Ouyang RY, Chen P. Epigenetic mechanisms in chronic obstructive pulmonary disease. Eur Rev Med Pharmacol Sci. 2015;19:844–856.
23. Huang T, Chen X, Hong Q, et al. Meta-analyses of gene methylation and smoking behavior in non-small cell lung cancer patients. Sci Rep. 2015;5 8897.
24. Grucza RA, Johnson EO, Krueger RF, et al. Incorporating age at onset of smoking into genetic models for nicotine dependence: evidence for interaction with multiple genes. Addict Biol. 2010;15:346–357.
25. Olfson E, Edenberg HJ, Nurnberger J Jr, et al. An ADH1B variant and peer drinking in progression to adolescent drinking milestones: evidence of a gene-by-environment interaction. Alcohol Clin Exp Res. 2014;38: 2541–2549.
26. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry 2011;168:1041–1049.
27. Angrist, J, Pischke J. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect. 2010;24:3–30.
28. Imbens G, Rubin D, Sacerdote B. Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. Am Econ Rev. 2001;91:778–794.
29. Sacerdote B. How large are the effects from changes in family environment? A study of Korean American adoptees. Q J Econ. 2007;122:119–157.

Cite this article as: Bierut L, Cesarini D (2015) How genetic and other biological factors interact with smoking decisions. *Big Data* 3:3, 198–202, DOI: 10.1089/big.2015.0013.

Abbreviations Used

GWAS = genome-wide association study
SNP = single nucleotide polymorphism
KHP = Kavli Human Project
DNA = deoxyribonucleic acid
CPD = cigarettes per day
TAG = Tobacco and Genetics Consortium
GxE = gene x environment

Big Data
Volume 3 Number 3, 2015
DOI: 10.1089/big.2015.0014

ORIGINAL ARTICLE

Diets and Health: How Food Decisions Are Shaped by Biology, Economics, Geography, and Social Interactions

Adam Drewnowski,[1,*] and Ichiro Kawachi[2]

Abstract

Health is shaped by both personal choices and features of the food environment. Food-choice decisions depend on complex interactions between biology and behavior, and are further modulated by the built environment and community structure. That lower-income families have lower-quality diets is well established. Yet, diet quality also varies across small geographic neighborhoods and can be influenced by transportation, retail, and ease of access to healthy foods, as well as by attitudes, beliefs, and social interactions. The learnings from the Seattle Obesity Study (SOS II) can be usefully applied to the much larger, more complex, and far more socially and ethnically diverse urban environment of New York City. The Kavli HUMAN Project (KHP) is ideally positioned to advance the understanding of health disparities by exploring the multiple underpinnings of food decision making. By combining geo-localized food shopping and consumption data with health behaviors, diet quality measures, and biomarkers, also coded by geographic location, the KHP will create the first-of-its-kind bio-behavioral, economic, and cultural atlas of diet quality and health for New York City.

Key words: food decisions; nutrition; economics; geography; behavior; public health

Introduction

The seminal work on *Food, Health, and Incomes* (1936) by John Boyd-Orr highlighted the public health problem of malnutrition among the poor in depression-era Great Britain.[1] Boyd-Orr was among the first to look at the socioeconomic gradient in diet quality, link it to food prices, and apply his findings to public policy. As the first Director-General of the Food and Agriculture Organization (FAO) of the United Nations, he worked for a more equitable and affordable global food supply.

The same socioeconomic gradients in diet quality and health can be seen in New York City today. However, the nature of malnutrition has changed. Excess empty calories of minimal nutritional value are now associated with higher rates of overweight, obesity, type 2 diabetes mellitus (T2DM), and metabolic syndrome (MetS). The relative prevalence of those diet-related diseases can be reliably mapped across diverse New York City neighborhoods.

The recently completed Seattle Obesity Study (SOS) used novel GIS/GPS techniques to bring inequalities in diets and health into sharp relief at the neighborhood level.[2] Obesity rates varied by as much as fivefold across Seattle neighborhoods, a range far greater than was observed with ethnicity or incomes. Analyses of data for >59,000 insured persons from a local HMO showed that obesity rates among women were predicted by house prices assessed at both tax parcel and at neighborhood level.[2]

Both in Seattle and nationally, minorities and lower-income groups had lower-quality diets and higher rates of obesity, T2DM, and MetS.[2–4] Diet quality could also be mapped across neighborhoods. In other studies, the

[1]*University of Washington, Seattle, Washington.*
[2]*Harvard University, Cambridge, Massachusetts.*

*Address correspondence to: *Adam Drewnowski, University of Washington, Seattle, WA*, E-mail: adrewnow@fredhutch.org

observed socioeconomic disparities in diets and health were related to retail food prices and diet costs.[5] Taken together, this body of work suggests that the multiple influences on food decision making, some economic and some not, strongly influence diet quality and affect likely health outcomes in the long term.

The principal culprits in today's obesity epidemic appear to be excess refined grains, added sugars, and fats. These foods are palatable (some say addictive), energy dense, inexpensive, culturally appropriate, and widely accessible. Even so, some population subgroups show what is called nutrition resilience—the ability to construct diets that are nutrient-rich, affordable, and appealing. How nutrition resilience is shaped by biology, psychology, or economics remains to be seen.

The psychosocial context of diet and disease is highly complex. Public health policy and practice, including the design of key dietary interventions, will benefit from a research focus on the neighborhood food environment and the behavioral economics of food choice. Multiple food-related decisions are made every day. Those decisions may be linked to the underlying neurobiology, but also to economics, culture, or the food environment. What is more, those decisions are influenced by social interactions, and they can vary from one neighborhood to another.

The exact mechanisms that underlie food purchase behaviors and so determine both diet quality and health are not always clear. There are complex interactions among social class, food and built environments, diet quality measures, and obesity rates. A better understanding is also needed of what local factors shape food decision making within the neighborhood context. The next phase of research on food-based decision making will require contributions from neuroscientists, nutritionists, epidemiologists, geographers, and urban planners.

A number of questions in multiple domains will need to be answered:

Question 1. Economic: do healthier diets cost more?

Empirically, the answer to the question "Do higher-quality diets generally cost more?" appears to be yes. Recent reviews and meta-analysis support that conclusion; on average, a healthier diet costs an additional $1.50 per person per day.[6] However, healthier diets would not cost more if different food-related decisions were made. Based on analyses of diet quality in relation to cost, it appears that some individuals and minority groups (e.g., Mexican Americans) are able to eat better for less. This phenomenon is called nutrition resilience, given its relation to optimal decision making in face of economic adversity. However, broader cultural acceptance of healthy yet inexpensive foods and the avoidance of calorie-dense packaged foods is one topic that needs further research.

The economic issues extend to food decision making by food assistance recipients. One question is whether Supplemental Nutrition Assistance Program (SNAP) recipients exhibit payday effects; for example eating lower-quality food toward the end of the month when the money runs out. An analysis of inpatient records in California suggests that hospital admissions for hypoglycemia increased 27% in the last week of the month for low-income patients with diabetes compared with the first week.[7] No such trend was observed among high-income individuals. Since many government checks arrive in the mail at the start of the month, changing the system to more frequent payments might help recipients smooth their consumption and improve their nutritional choices. Such payments may include Social Security, SNAP, and Temporary Assistance for Needy Families (TANF). More broadly, research is needed to understand the potential effects of regular food insecurity on stress response and health.

Question 2. Environmental: should access to healthy foods continue to be measured in terms of physical distance?

These issues relate to the distribution of the food supply in urban centers such as New York City. The relevant question is "Do residents of low-income/minority neighborhoods in New York City face poorer food choices?" Research shows that food choices in a given neighborhood depend directly on the purchasing power of that neighborhood. Access to healthy foods may be largely economic in nature and have less to do with physical distance. Building new supermarkets in low-income neighborhoods may not improve nutritional status in the absence of economic interventions. Build it and they may not come.

A key topic of research in the KHP is to understand better where low-income residents shop for food and what compromises they make. This is related to the extent to which "food deserts" really matter. In the SOS, it was found that people did not shop at the nearest supermarket—rather, they selected one within their price range.[8] One thought is that food access ought to be measured in terms of economic access. In the New York City

context, transport will make a difference, since many people do not have cars and do not load a week's worth of groceries into the trunk.

Question 3. Eating in or eating out?

One common idea is that "time poverty" forces low-income families to rely on prepared packaged foods or source meals from fast-food restaurants. Seattle-based studies show that the frequency of home-cooked meals and the time spent on food preparation, cooking, and cleaning were all linked to higher-quality diets. More likely to eat away from home were young working adults and single-parent families, who placed a greater premium on time and convenience.[9] Preliminary analyses of the nationally representative National Health and Nutrition Examination Survey (NHANES) suggest that most likely to cook at home were large families, Latinos, and groups of lower education and incomes. Again, food decisions may be influenced by ethnicity and culture. For example, the traditional Mexican American diet can be prepared from relatively inexpensive ingredients, and traditions of food preparation may vary across social groups. Another key issue here is the nature of the family unit. To what extent do households in which several generations either share the same roof or live very close together share food preparation? In such cases, do members of the older generations lighten the load on younger members of the family who are working?

Question 4. Psychological: Do nutrition-related attitudes count?

An important and unanswered question is whether low socioeconomic status (SES) shoppers are more susceptible to advertising claims on packaged junk foods such as "organic" or "gluten-free." Generally speaking, high SES shoppers are more likely to be "label literate," that is, to understand and act on information provided to consumers. Another question concerns the extent to which those with chronic conditions, such as diabetes and obesity, shop in a health-conscious way, selecting zero-calorie or lower-calorie beverages and diet foods. Also of interest is how this relationship itself is moderated by SES. The value of the KHP in this regard is clear, since there are limited data on this topic. In the SOS, survey questions based on the NHANES consumer module revealed that positive attitudes toward nutrition improved diet quality. The fact that this appeared to be true at every level of education and income suggests that this may not contribute to inequality-related diet quality.[10] However, it is a question that is profound interest for its own sake and that remains little understood.

Question 5. Biological: what are the biomarkers of behavior?

KHP will make it possible to study the biological-mediating mechanism in the pathway between SES and health outcomes. Researchers have focused on stress and response to stress.[11–13] Stressors can be personal or environmental (traffic noise, pollution, crime). The biomarkers that have been studied include cortisol and inflammation biomarkers such as c-reactive protein. At least some of these biomarkers have been related to SES. Telomeres, a marker of accelerated aging, have also been tied to SES, stress, and low-quality diets.[11–13]

An important tool in understanding the above mechanisms will be a socioeconomic and bio-behavioral diet and health atlas for New York City. By combining biomarkers data with geo-localization techniques, the KHP will be well positioned to create the first-of-its-kind such atlas. In this respect, the KHP can build on the insights identified in the SOS. As indicated, the SOS explored links between multiple characteristics of neighborhoods, such as SES, social capital, physical and economic access to food sources, and opportunities for physical activity and obesity rates. The neighborhood variables were both real (assessed by GPS/GIS methods) and perceived (assessed by questionnaire self-report). The SOS helped to advance obesity research by transforming geographic and economic data into individual-level variables for use in studies on diets, health, and weight. Importantly, it distinguished between physical access to food sources, measured in terms of distance, and economic access to foods, measured in terms of food prices and diet costs. Perceived physical access was measured in terms of perceived distance and/or length of travel to principal food sources, including supermarkets, convenience stores, and fast-food outlets. Economic access was measured in terms of perceived food expenditures per week at various shopping and eating locations. The SOS team has devised methods to assess the price of a market basket across supermarket chains, as well as new procedures to estimate diet cost, real and perceived. Perceived expenditures were validated using actual expenditures backed by 2-week receipts for all foods at home and away from home.

One important focus of the SOS was on food retail and food shopping decisions. Obese shoppers were much more likely to shop in lower-cost grocery stores; the prevalence of obesity varied dramatically by store type.[8] Shoppers at the lowest-price stores consumed fewer fruit and vegetable servings compared with those shopping at the highest-price stores and were more likely to be obese (body mass index >30 kg/m^2). Few people shopped for food in their residential neighborhood or home census tract. In Seattle, distance to the food destination varied by race/ethnicity, income, and education. Whites were more likely to shop at the closest supermarket (1.5 miles), whereas African Americans showed the largest disparity between the closest supermarket and the one used (3 miles).

Implementation in the KHP

Investigation of factors that contribute to food decisions would utilize the following KHP data sets, amongst others: (a) Dietary data would be collected via periodic food diaries, complemented by mining for food purchases in financial data (credit cards, debit cards, checks). (b) Data on decisions to eat in or eat out and physical distance to different food sources would be available via geo-location data from smartphones and activity trackers, mapped to GIS. (c) Information on financial status and participation in government assistance programs (SNAP, Social Security, TANF) would be available via financial data gathered using a combination of automated and survey-based methodologies. (d) Attitudes toward healthy foods would be assessed via questionnaires on smartphones or tablets. (e) Levels of "life stress" would be measured via self-administered psychological questionnaires on digital devices. In addition, cortisol levels in saliva would be measured triennially, starting at study intake. (f) Demographics and race/ethnicity data would be available via the KHP questionnaire at study intake.

The impact of food decisions on health would be analyzed via the following KHP data sets, amongst others: (a) Medical information on study participants' health would be available from the medical history and records going forward (EMRs, doctors' notes, hospital records, dental records). Prescription data would be gathered via the New York State Prescription database. This information would be complemented by the SPARCS database and the KHP's own tests: blood tests (blood metals, vitamins, lipids, glucose, and other biomarkers), urine tests, and hair tests (smoking, alcohol, and substance use) every 3 years. (b) Information on genetic variation at the individual and family levels would be gathered via whole genome sequencing of blood samples for adults (saliva for children) performed at study intake. In addition, data on variation in epigenetics would be gathered via triennially performed assays. (c) Cognitive function would be measured via self-administered psychology questionnaires/tests on smartphones and tablets at intake and periodically thereafter.

Conclusion

The importance of decision-making processes in food selection cannot be overstated. Food purchase decisions and hence diet quality depend on a host of environmental factors, some modifiable and some not. All of those factors need to be studied in their neighborhood context. Identifying the key biological, economic, and environmental influences on these complex processes is of immense importance to public health.

Author Disclosure Statement

No competing financial interests exist.

References

1. Boyd-Orr J. Food Health and Income. London, Macmillan and sons, 1936, pp. 1–72.
2. Rehm CD, Moudon AV, Hurvitz PM, et al. Residential property values are associated with obesity among women in King County, WA, USA. Soc Sci Med. 2012;75:491–495.
3. Drewnowski A, Specter SE. Poverty and obesity: the role of energy density and energy costs. Am J Clin Nutr. 2004;79:6–16.
4. Rehm CD, Monsivais P, Drewnowski A. Relation between diet cost and Healthy Eating Index 2010 scores among adults in the United States 2007–2010. Prev Med. 2015;73:70–75.
5. Darmon N, Drewnowski A. Does social class predict diet quality? Am J Clin Nutr. 2008;87:1107–1117.
6. Rao M, Afshin A, Singh G, et al. Do healthier foods and diet patterns cost more than less healthy options? A systematic review and meta-analysis. BMJ Open. 2013;3:e004277.
7. Seligman HK, Bolger AF, Guzman D, et al. Exhaustion of food budgets at month's end and hospital admissions for hypoglycemia. Health Aff (Millwood). 2014;33:116–123.
8. Drewnowski A, Aggarwal A, Hurvitz PM, et al. Obesity and supermarket access: proximity or price? Am J Public Health 2012;102:e74–80.
9. Monsivais P, Aggarwal A, Drewnowski A. Time spent on home food preparation and indicators of healthy eating. Am J Prev Med. 2014;47:796–802.
10. Aggarwal A, Monsivais P, Cook AJ, et al. Positive attitude toward healthy eating predicts higher diet quality at all cost levels of supermarkets. J Acad Nutr Diet. 2014;114:266–272.

11. Geronimus AT, Pearson JA, Linnenbringer E, et al. Race-ethnicity, poverty, urban stressors, and telomere length in a Detroit community-based sample. J Health Soc Behav. 2015;56:199–224.
12. Needham BL, Carroll JE, Diez Roux AV, et al. Neighborhood characteristics and leukocyte telomere length: the multi-ethnic study of atherosclerosis. Health Place 2014;28:167–172.
13. Park M, Verhoeven JE, Cuijpers P, et al. Where you live may make you old: the association between perceived poor neighborhood quality and leukocyte telomere length. PLoS One 2015;10:e0128460.

Cite this article as: Drewnowski A, Kawachi I (2015) Diets and health: how food decisions are shaped by biology, economics, geography, and social interactions. *Big Data* 3:3, 193–197, DOI: 10.1089/big.2015.0014.

Abbreviations Used

FAO = Food and Agriculture Organization
T2DM = type 2 diabetes mellitus
MetS = metabolic syndrome
SOS = Seattle Obesity Study
SNAP = Supplemental Nutrition Assistance Program
TANF = Temporary Assistance for Needy Families
NHANES = National Health and Nutrition Examination Survey
SES = socioeconomic status
SPARCS = Statewide Planning and Research Cooperative System
EMRs = electronic medical records

About:

What are the roots of human behavior? How do we make decisions, and what forces shape those decisions? How can we apply what we learn about human health and behavior to inform and improve public policy to make it more effective? The answers to these questions lie in an entirely new way of studying human beings, taking advantage of technological advances and the big data revolution to provide a previously unattainable interdisciplinary data platform to researchers.

Traditional longitudinal studies have been focused on specific domains of inquiry or subsets of the population. In contrast, the Kavli HUMAN Project is the next step in bio-behavioral and longitudinal research by measuring all of the biological and behavioral characteristics that make us human at once, effectively quantifying the human condition.

The Kavli HUMAN Project's data platform will provide researchers with an unprecedented volume and diversity of datatypes to break down the silos between disciplines like neuroscience, genetics, psychology, medicine, and urban informatics in order to unlock new insights into the feedback mechanisms between biology, behavior, and the environment that make up the bio-behavioral complex. Over time, the Kavli HUMAN Project will not only enable groundbreaking science, but provide the context and knowledge to craft evidence-based public policies and improve societal outcomes.