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.

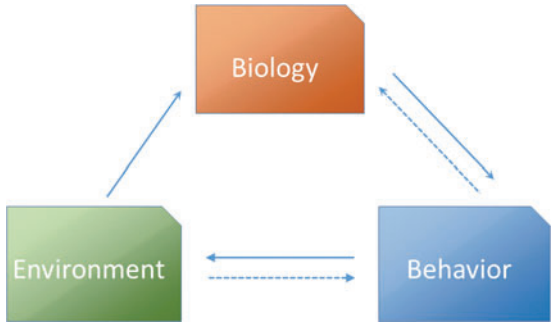

FIG. 1. The Biobehavioral Complex.

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.

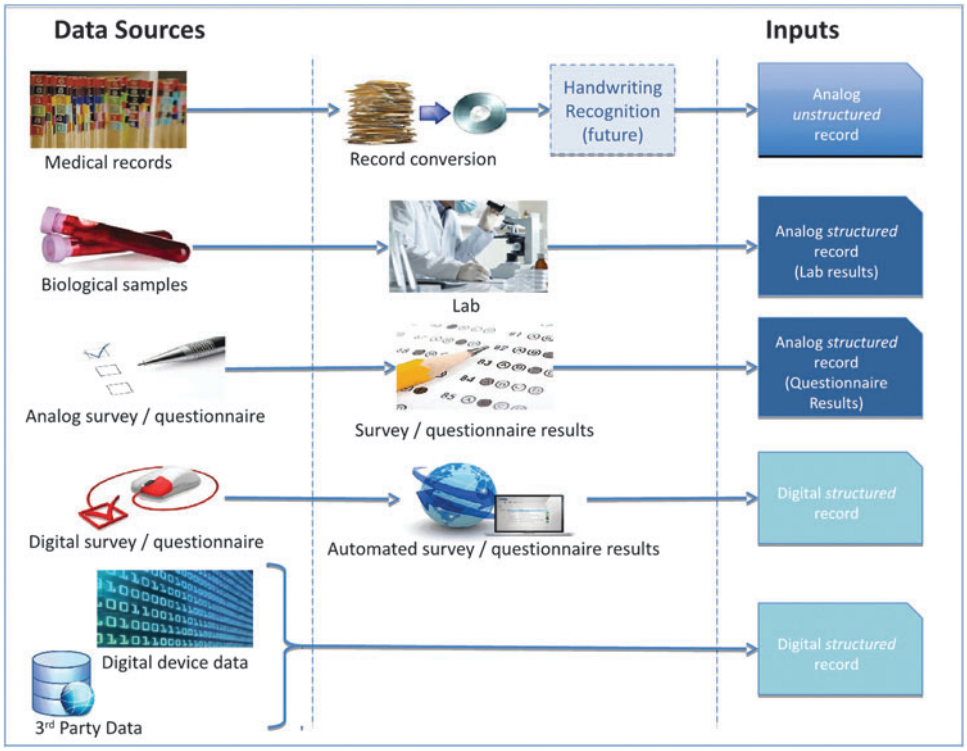

FIG. 2. Data sources to inputs for the KHP study.

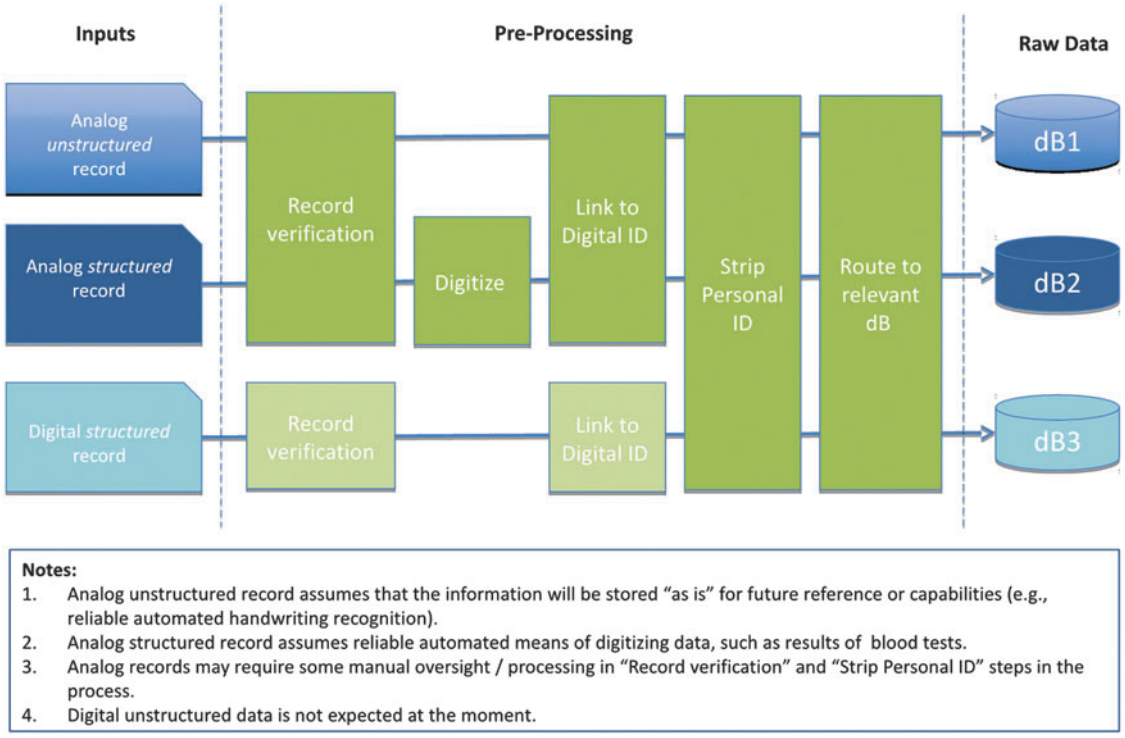

FIG. 3. Data ingestion for KHP study.

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