Personal Data for the Public Good

New Opportunities to Enrich Understanding of Individual and Population Health

FINAL REPORT OF THE HEALTH DATA EXPLORATION PROJECT
MARCH 2014

Health Data Exploration project

Supported by

Robert Wood Johnson Foundation

Conducted by the California Institute for Telecommunications and Information Technology

UC Irvine

UC San Diego

Contents


WeightWatchers

1. Executive Summary

Individuals are tracking a variety of health-related data via a growing number of wearable devices and smartphone apps. More and more data relevant to health are also being captured passively as people communicate with one another on social networks, shop, work, or do any number of activities that leave "digital footprints."

Almost all of these forms of "personal health data" (PHD) are outside of the mainstream of traditional health care, public health or health research. Medical, behavioral, social and public health research still largely rely on traditional sources of health data such as those collected in clinical trials, sifting through electronic medical records, or conducting periodic surveys.

Self-tracking data can provide better measures of everyday behavior and lifestyle and can fill in gaps in more traditional clinical data collection, giving us a more complete picture of health. With support from the Robert Wood Johnson Foundation, the Health Data Exploration (HDE) project conducted a study to better understand the barriers to using personal health data in research from the individuals who *track* the data about their own personal health, the companies that market self-tracking devices, apps or services and *aggregate and manage* that data, and the researchers who might *use* the data as part of their research.

Perspectives

Through a series of interviews and surveys, we discovered strong interest in contributing and using PHD for research. It should be noted that, because our goal was to access individuals and researchers who are already generating or using digital self-tracking data, there was some bias in our survey findings—participants tended to have more education and higher household incomes than the general population. Our survey also drew slightly more white and Asian participants and more female participants than in the general population.

Individuals were very willing to share their self-tracking data for research, in particular if they knew the data would advance knowledge in the fields related to PHD such as public health, health care, computer science and social and behavioral science. Most expressed an explicit desire to have their information shared anonymously and we discovered a wide range of thoughts and concerns regarding thoughts over privacy.

Equally, researchers were generally enthusiastic about the potential for using self-tracking data in their research. Researchers see value in these kinds of data and think these data can answer important research questions. Many consider it to be of equal quality and importance to data from existing high quality clinical or public health data sources.

Companies operating in this space noted that advancing research was a worthy goal but not their primary business concern. Many companies expressed interest in research conducted outside of their company that would validate the utility of their device or application but noted the critical importance of maintaining their customer relationships. A number were open to data sharing with academics but noted the slow pace and administrative burden of working with universities as a challenge.

In addition to this considerable enthusiasm, it seems a new PHD research ecosystem may well be emerging. Forty-six percent of the researchers who participated in the study have already used self-tracking data in their research, and 23 percent of the researchers have already collaborated with application, device, or social media companies.

The Personal Health Data Research Ecosystem

A great deal of experimentation with PHD is taking place. Some individuals are experimenting with personal data stores or sharing their data directly with researchers in a small set of clinical experiments. Some researchers have secured one-off access to unique data sets for analysis. A small number of companies, primarily those with more of a health research focus, are working with others to develop data commons to regularize data sharing with the public and researchers.

> *SmallStepsLab serves as an intermediary between Fitbit, a data rich company, and academic researchers via a "preferred status" API held by the company. Researchers pay SmallStepsLab for this access as well as other enhancements that they might want.*

These promising early examples foreshadow a much larger set of activities with the potential to transform how research is conducted in medicine, public health and the social and behavioral sciences.

Opportunities and Obstacles

There is still work to be done to enhance the potential to generate knowledge out of personal health data:

- **Privacy and Data Ownership:** Among individuals surveyed, the dominant condition (57%) for making their PHD available for research was an assurance of privacy for their data, and over 90% of respondents said that it was important that the data be anonymous. Further, while some didn't care who owned the data they generate, a clear majority wanted to own or at least share ownership of the data with the company that collected it.

- **Informed Consent:** Researchers are concerned about the privacy of PHD as well as respecting the rights of those who provide it. For most of our researchers, this came down to a straightforward question of whether there is informed consent. Our research found that current methods of informed consent are challenged by the ways PHD are being used and reused in research. A variety of new approaches to informed consent are being evaluated and this area is ripe for guidance to assure optimal outcomes for all stakeholders.

- **Data Sharing and Access:** Among individuals, there is growing interest in, as well as willingness and opportunity to, share personal health data with others. People now share these data with others with similar medical conditions in online groups like PatientsLikeMe or Crohnology, with the intention to learn as much as possible about mutual health concerns. Looking across our data, we find that individuals' willingness to share is dependent on what data is shared, how the data will be used, who will have access to the data and when, what regulations and legal protections are in place, and the level of compensation or benefit (both personal and public).

- **Data Quality:** Researchers highlighted concerns about the validity of PHD and lack of standardization of devices. While some of this may be addressed as the consumer health device, apps and services market matures, reaching the optimal outcome for researchers might benefit from strategic engagement of important stakeholder groups.

We are reaching a tipping point. More and more people are tracking their health, and there is a growing number of tracking apps and devices on the market with many more in development. There is overwhelming enthusiasm from individuals and researchers to use this data to better understand health. To maximize personal data for the public good, we must develop creative solutions that allow individual rights to be respected while providing access to high-quality and relevant PHD for research, that balance open science with intellectual property, and that enable productive and mutually beneficial collaborations between the private sector and the academic research community.

"I'm happy to contribute [my data] if it could contribute to, say, a larger study where there could be some additional knowledge."

– Individual

"One of the main strengths of this research is that it has potential to be very translational. A lot of the findings that can come out of it can be directly applied in people's lives and are related to the types of health outcomes that people care about a lot."

– Researcher

"If anything, having research institute academically published on some of the data would help give us more credibility in the market. From a company we are interested in it."

– Company

2. Introduction

A variety of health-relevant parameters are now being captured via an ecosystem of consumer-oriented wearable devices, smartphone apps and related services. Ever larger streams of data are being produced by individuals—across lifespans, throughout the course of health and illness and in geospatial context. In early 2013, the Pew Foundation's Tracking for Health study found that 69% of Americans track some form of health related information and fully 21% of them use some form of digital device to do so (Fox & Duggan, 2013). An indicator of the momentum behind the trend to produce and collect data about ourselves, or self tracking, may be seen in a small but growing Quantified Self movement, in which individuals meet together to share insights they have gained from their self tracking activities. There is growing interest in, as well as willingness and opportunity to, share personal health data with others. People now share these data with others with similar medical conditions in online groups like PatientsLikeMe or Crohnology, with the intention to learn as much as possible about shared health concerns. The trend for sharing extends to opening up personal health data to see what insights others might see in them.

In addition to self-tracked and voluntarily shared personal health data, more and more data about individuals is being captured passively as people communicate with one another on social networks, shop, work, or do any number of activities that leave "digital footprints" in the increasingly expanding "Internet of Services." Industry has capitalized on this trend to refine and personalize services and marketing, often to a remarkable degree (Turow, 2011).

Almost all of these forms of data, herein denoted as "personal health data" (PHD) (Clarke et al., 2007), share one thing: the devices, apps and service that capture and store them are owned by entities that are outside of the mainstream of traditional health care, public health or health research. This includes everything from small start-ups to globally active consumer electronic, telecommunications, computer and social network corporations.

At the same time, medical, behavioral, social and public health research still largely rely on traditional sources of health data such as those collected in clinical trials funded by the pharmaceutical industry or the National Institutes of Health, sifting through electronic medical records, or conducting periodic surveys of representative samples of individuals to make inferences about broader behavioral, social or public health trends. The quality of data collected through these methods may be high, but this comes with a cost, including how much and how frequently these data can be collected. Also, almost by definition these traditional methods of health research can't capture the multidimensional and continuous nature of the behavioral, social and environmental influences that are increasingly recognized as critical to human health (Glass & McAtee, 2006).

With this as background, in mid-2013, The Robert Wood Johnson Foundation funded the Health Data Exploration (HDE) project to gain further insights into how various stakeholder groups think about personal health data and its use for research. Stakeholders include: a) individuals who self-track and/or share health-related data; b) health researchers with an interest in how to use these new forms of data to gain insights into personal and population health; c) the companies that market the devices, applications and services that generate these data; and d) key informants from the worlds of health care, public health and health policy. This report provides the results of this effort. At a high level, our goal is to Identify barriers and opportunities to unconvering new health insights from these kinds of data.

The HDE project began with the development of an advisory board of thought leaders in the areas most relevant to this project. Advisors were asked to share their insights about the both existing and emerging trends in these new forms of health data. To further inform the project, an environmental scan was conducted to identify peer-reviewed and other scientific publications, foundation reports, governmental reports, key thought pieces in the popular media and other sources. This led

to both a research synthesis (Section 3) and an Annotated Bibliography (Section 7; Appendices) that should be of value to individuals, companies, researchers and policy makers interested in this space.

These efforts overlapped with the deployment of an online survey, conducted from August 1, 2013 to September 11, 2013, of individuals and researchers. The methods and key results from this survey are presented in Section 4. In addition to the survey, in depth interviews were conducted with representatives from each of the three stakeholder groups as well as key informants to develop a deeper understanding of the issues that surfaced in the surveys as well as themes discovered in our discussions with advisory board members and the literature review. Several vignettes of the findings of these interviews are also presented in Section 4.

Throughout this process, several key issues emerged that required detailed analysis and discussion. Many of these issue cluster around the importance of trust in establishing the ecosystem that will support individuals donating their data for public research. Specific issues include privacy related to personal data (Section 5.1), human subjects research and informed consent (Section 5.2) and data sharing (Section 5.3). Each of these sections describes what was learned from our interviews, from discussions with the advisory board members and key informants, and through a review of the literature. Since each area is worthy of a full-length monograph in itself, what is provided here is only an overview of the issues.

Finally, based upon this background, several opportunities and obstacles related to progress in the field of personal health data research are briefly discussed (Section 6).

3. Background

3.1 New Devices, New Data

Given their growing ubiquity, smart phones and wearable devices have gained the attention of researchers, marketers and app makers. Applications for sensing, storing and inputting health and activity data have proliferated, and are increasingly being used by a wide range of individuals for self-tracking. The usefulness of smart phones and other devices for collecting data can be expected to increase with the continued miniaturization of sensors and other embedded technologies (Davies, 2013). Health and lifestyle data is abundantly produced and collected in the ordinary course of daily life for many people. Additionally, consumers are now able to directly purchase sophisticated tests, including blood tests and direct-to-consumer genetic tests, adding to stores of "big data" with potential for public health research.

In related technological trends, computing and storage technologies have decreased in price and sensing and networking infrastructures have sufficiently developed that we are dealing with a "data deluge" in multiple research domains (Borgman, Wallis, & Enyedy, 2007). Environmental and other sciences are struggling to develop and implement consistent best practices so that data can be obtained and stored in a way that maximizes utility and re-use (Edwards et al., 2013). Research methods for making use of "big data" are being developed as researchers envision the potential for novel way to analyze complex phenomena.

3.2 New Opportunities in Research

From this combined technological and social state of affairs, several opportunities for public health research have emerged. First, the plethora of apps and devices that are commercially available both allow and entice people to easily collect, store, and analyze data about their ordinary behaviors and activities, and encourages them to use that data to intervene in those behaviors and activities. In turn, people may participate in online communities devoted to sharing health and disease experience and self-tracking data, or even join the Quantified Self movement, tracking many aspects of their biology and health, taking genetics tests and sharing this information amongst participants and with researchers. The "formation of new group and individual identities and practices" in response to these trends in data collecting and sharing has been termed "biosociality" (Rabinow, 1999).

The Quantified Self movement promises "self knowledge through numbers" and its adherents are proponents of self-tracking in many forms, including the use of wearable devices, blood testing, genetic testing, and journaling. Self quantifiers track activity, diet, mood, sleep, and as many other parameters as possible. Participants iterate through stages including collection, reflection and action (Li, Dey, & Forlizzi, 2010) and seek to answer questions regarding status, history and goals (Li, Dey, & Forlizzi, 2011). They may also meet in groups or use Internet discussion boards to share experiences and compare findings.

In addition to social trends that accompany self-tracking technologies, opportunities to develop novel research methods and projects have emerged along with these prolific new data sources. The analysis of person-generated data has been called "reality mining" and can be applied in issues of individual health, social networks, behavioral patterns, infectious disease and mental health (Pentland, Lazer, Brewer, & Heibeck, 2009). For example, Internet discussion forums can be mined for evidence about improperly functioning lens implants (Hagan & Kutryb, 2009). Ayers and co-authors developed methods for linking internet searches to economic indicators to gauge population distress in real time, rather than retrospectively, and for analyzing Google queries to monitor seasonal changes in mental health at the population level (Ayers et al., 2012; Ayers, Althouse, Allem, Rosenquist, & Ford, 2013). Data generated as byproducts of daily life can be predictive of social behaviors, for example shopping (Krumme, Llorente, Cebrian, Pentland, & Moro, 2013) and location (Song, Qu, Blumm, & Barabási, 2010). These technologies can be used to model and

predict human behavior (Lane et al., 2011). Researchers used anonymized cell phone data from 100,000 users to characterize individual travel patterns (González, Hidalgo, & Barabási, 2008). Lane et al. (Lane et al., 2010) describe existing sensor technologies in smart phones and propose a framework for future research that makes use of the dispersion of these technologies.

Self-tracking and device data have potential for a range of public health inquiries, including epidemiology and mental health. Researchers used specialized software on mobile phones to identify peer interactions and track characteristics including cold/flu state, mental health, and obesity status (Madan, Cebrian, Lazer, & Pentland, 2010; Madan, Cebrian, Moturu, Farrahi, & Pentland, 2012). Unhealthy eating and exercise levels could also be detected (Madan, Moturu, Lazer, & Pentland, 2010). Data can be used to provide objective measures for tracking depression (Sung, Marc, & Pentland, 2005). The relationship between sleep and mood has also been explored using cell phone and Bluetooth data combined with self reports (Moturu, Khayal, Aharony, Pan, & Pentland, 2011). One study used Fitbit devices to count steps of patients recovering from surgery, finding that the more steps walked, the shorter the hospital stay and the less likely patients would need care in a nursing facility (Cook, Thompson, Prinsen, Dearani, & Deschamps, 2013). Wearable devices can aid weight loss goals as much as support groups (Pellegrini et al., 2012).

In addition to using device data for research, the potential for genetic repositories has been explored using data from 23andMe and the Personal Genome Project. Researchers identified two genetic associations for Parkinson's disease using 23andMe genetic data and self-reports (Do et al., 2011). These data have also been used to identified genes for traits such as freckling, curly hair, and photic sneezing (Eriksson et al., 2010). Using cell lines from an individual donor to the Personal Genome Project, authors characterized allele-specific DNA methylation and its role in fuzzy methylation (Shoemaker, Deng, Wang, & Zhang, 2010). Researchers developed an RNA-guided genome editing system and used Personal Genome Project data to create a "genome-wide reference of potential target sites in the human genome" (Mali et al., 2013). Researchers and funding agencies like NIH and NSF are seeking new ways to extract medically and biologically relevant information from datasets and provide access to publicly produced datasets. An example of this is the 1000 Genomes Project and its partnership with private companies like Amazon Web Services (Conger, 2012).

An opportunity presented by the growing amount of PHD may be to move beyond the use population-level data for simple descriptive epidemiology to its use to infer causality. Fundamental principles of epidemiology are based upon how causality should be determined (Hill, 1965). These were developed at a time when health-related measures were usually infrequently collected and expensive in time, materials and participant burden. These barriers are now often dramatically reduced by the increasing ubiquity of PHD. It is possible now that we may have sufficient data on a variety of determinants of health that we may be on the cusp of a new form of establishing causality, akin to how researchers in fields like atmospheric science or economics make predictions about future events from the models they develop on ever-changing real time data sets.

3.3 New Challenges

These new methods of acquiring data and approaching research have raised new challenges with familiar issues. Three areas of interest are privacy, consent and data access.

Privacy norms and expectations are becoming more diverse, stretched in opposite directions by opposing trends. On the one hand, there is increased sharing in an era of online communication and social networking sites like Facebook, Twitter, and Tumblr. Only a small percentage of college students change their privacy setting (Gross & Acquisti, 2005). The "born digital" generation has different expectations of privacy, increasing social pressure to share, and entire lives documented in online content (Palfrey & Gasser, 2008), and some of these are racially differentiated (Madden et al., 2013).

On the other hand, there is increased desire for privacy in response to adverse events. 55% of surveyed Internet users have taken steps to avoid observation by specific people, organizations, or the government. 6% of those surveyed reported having their reputation damaged by online activity. (Raine et al, 2013). Publically available genetic data that was thought to be

properly anonymized was shown to be vulnerable to de-anonymization (Gymrek, McGuire, Golan, Halperin, & Erlich, 2013; Homer et al., 2008). This lead to the removal of public access and calls for a re-evaluation of the role of IRBs in light of new research methods and data sources (Lazer et al., 2009).

Closely related to privacy is the need for informed consent. The case of Henrietta Lack has drawn popular attention to the problem of botched informed consent and raised the question of family's rights when shared genetic information is made public (Ahmed, 2013; Zimmer, 2013). When publishing an article based on data from 23andMe, the editors of Public Library of Science (PLoS) explained their concerns about the lack of informed consent data before publishing research based on data from 23andMe (Gibson & Copenhaver, 2010). The work was not classified as "human subjects research" because it did not meet either criteria of (a) the researchers obtaining data directly from subjects or (b) the researchers being able to identify the subject. However, they noted that informed consent would have been ideal and that there was a need for clear policies in this new gray area.

Data access becomes more complicated when researcher acquire data from companies rather than collect it directly. Whereas big data technologies in physics and genomics were heavily developed by academics and funded by universities or public agencies, many of the resources relevant to Health Data Exploration are commercially developed. Datasets can be proprietary or have significant strategic value. Research based on privately share data has raised concerns about verification and reproducibility of the science, as well as the privileging of a few researchers with access to the data (Huberman, 2012). Additionally, norms for sharing data from publicly funded research are jeopardized by keeping these repositories of data private (Markoff, 2012). Some industry leaders and researchers have even argued that universities are no longer the most apt sites for medical and genetic research, but rather, private firms whose users generate massive quantities of data, like Amazon.com and Facebook, (Markoff, 2012).

Even when data do not have proprietary restrictions, there is the potential for researchers to improve data sharing practices. A review of thousands of previously published phylogenetic studies estimated that two-thirds of the studies did not make any data available beyond the article figures (Drew et al., 2013). As data sets grow, there are more opportunities for exploration beyond the original intended use of the data, and lack of access prevents this reuse.

Public health research will inherit some of the same challenges as other "big data" projects but with several unique problems to solve—and opportunities to address—as well (Lazer et al., 2009). These include potential concerns that access to newer forms of low cost, easily accessible data as a potential substitute for population-level surveillance of public health issues will violate the privacy of citizens. An example of this can be seen with surveillance of dietary behaviors. Current methods use periodic sampling surveys such as the Behavioral Risk Factor Surveillance System that target respondents who are willing to answer a set of questions related to dietary behaviors. Measurement approaches based upon loyalty card data on food purchases from grocery consumers have demonstrated potential to expose important trends in diet patterns (Niederdeppe & Frosch, 2009). However, will these methods raise concerns about whether "big brother" is looking over our shoulder as we go about our daily lives?

3.4 This Project

Based upon this background, there is a need to better understand this new world of personal health data and its implications for improving personal and population health. The perspective of this project was not that these data would *supplant* current data-intensive efforts to understand health. Rather, the premise was that a better understanding of these new forms of data could potentially *complement and add value* to existing medical and public health efforts to measure the environmental, social, behavioral and medical determinants that comprise the full picture of health and society.

4. Surveys and Interviews

Given the need for a better understanding of the ecology of personal health data, we sought to elicit the experiences, behaviors and attitudes of three relevant stakeholder groups:

- **Individuals:** People who track data about their own personal health, including behaviors, metrics, and symptoms.
- **Researchers:** Researchers who may want to use self-tracking data as part of their research.
- **Companies and Key Informants:** Corporations that market self-tracking devices, apps or services, and companies that collect data on individuals that can provide insight into health-related states or events. Also included in this group are several key informants with specialized knowledge in personal health data research.

These three groups represent the primary stakeholders on the pathway from personal health data to public good research: the Individuals who *produce* the data, the Companies that *aggregate and manage* that data, and the Researchers who will *use* the data to produce research results.

We collected data using both survey and interview methods. For Individuals, both the survey and the interviews were aimed at understanding users' experiences with health tracking, the kinds of data they track, and their attitudes toward data sharing and privacy. For Researchers, our focus was on understanding the kinds of data that would be useful in various research domains, researchers' concerns about data quality and reliability, and their perception of barriers to the use of self-tracking data for research. For Companies, we conducted interviews with CEOs, technical managers, or other key employees to understand what data are collected, the legal, policy, and business concerns around these data, and companies' overall willingness and ability to make their data available to external researchers.

4.1 Survey Method

SURVEY DEVELOPMENT

We developed surveys to understand attitudes and experiences with self-tracking data for both Individuals and Researchers. Survey instruments were developed based on a set of high-level research questions developed by the research team. Questionnaires were pilot-tested and reviewed by experts before deployment. The high-level questions and full survey instruments are included in the Appendix. Surveys were administered using a local installation of LimeSurvey, an open-source survey management platform.

SAMPLING AND SURVEY DISTRIBUTION

A goal in our sampling was to access individuals and researchers who are already generating or using digital self-tracking data. The Pew Research Center's September 2012 Health Tracking survey found the only a relatively small segment of the population uses technology for self-tracking. Similarly, while some researchers are beginning to use self-tracking data in academic settings, these are still considered non-traditional data sources. Given the low percentage of early adopters in a general population, we chose to recruit participants through postings on related web pages, UCSD press releases, and various social-media channels including blogs and Tweets. The result is a targeted, self-selected sample.

In order to address the potential biases this sampling strategy produced in our survey, we asked a number of demographic questions that provide for comparisons to the general population. We also included some general questions that had been asked in the Pew Health Tracking Survey in order to calibrate our sample against Pew's national sample.

SURVEY ADMINISTRATION

The surveys were opened on August 1, 2013. The surveys were accessible through any web browser on an Internet-connected device. The surveys were closed on September 11, 2013.

As an incentive to participate in the surveys, participants who completed the survey were given the option to enter into a drawing for an iPad or Android tablet.

Table 1. **Number of survey participants**

Survey	Partial	Completed	Total
Individuals	104	361	465
Researchers	35	99	134

4.2 Interview Method

PROTOCOL DEVELOPMENT

We developed a separate interview protocol for each of the three groups: Individuals, Researchers, and Company/Key Informants. Interviews with Individuals and Researchers were designed to complement our survey by providing richness to the survey findings and eliciting data that would be difficult to collect in a survey. Company/Key Informants interviews included representatives of companies that provide personal health devices, apps, or services, as well as other experts in the area of personal health data. These interviews were designed to provide a map of corporations and other organizations active in the personal health data arena. For companies, we wanted to gauge their willingness to participate in collaborations with academic researchers and understand the business, technological, and social factors that affect their decision-making. We developed semi-structured interview protocols based around the same set of high-level questions that drove our initial survey design. We also drew on preliminary analyses of the survey data, identifying topics and questions with surprising or confusing results as candidates for further investigation.

INTERVIEW SAMPLING AND PROCEDURE

At the end of the surveys for Individuals and Researchers, we asked participants if they would be willing to be contacted to participate in follow-up interviews. We drew participants from this list. Individuals were chosen randomly, but stratified to ensure gender balance based on participant names. Researchers were chosen randomly, but stratified to ensure a balance of research interests. Participants were invited to participate by e-mail or telephone. Interviews were conducted in person or over the phone. Interviews were audio-recorded and transcribed.

For the Company and Key Informant interviews, targets were identified by the study team in collaboration with RWJF as well as based upon the advice of advisory board members. Detailed notes were taken for Company interviews to avoid confidentiality concerns associated with audio recording. We conducted a total of 35 interviews, including 11 individuals, 9 researchers, and 15 companies/key informants.

4.3 Survey Results

INDIVIDUALS

The individual survey was taken by 465 participants. Because we used a convenience sample, it is especially important to investigate the sampling bias in our survey. In order to provide a baseline, we compared the demographic characteristics of our population to known population statistics and the sample in the Pew Research Center's September 2012 Health Tracking survey. Overall, our survey tends to include more female participants (65%) than male (35%). Compared to the 2010 U.S. census, our survey also drew slightly more white and Asian participants than in the general population (Figure 1), and fewer Hispanic participants (3.8% vs. 16.4% in the U.S. population). Our sample also had a higher level of education than is found in general population surveys, with 90.4% of our sample having a 4-year college degree or higher. Our participants also tend to have higher incomes than the general population, with 47% of our participants in households with annual incomes of more than $100,000 per year.

Our survey participants, as expected, are primarily people who keep track of their personal health data. In our sample, 91% report tracking personal health data for themselves or a loved one, while only 69% of the participants in the Pew survey do. Pew also found that only 21% of U.S. adults use some form of technology to track their health data, while 65% of our sample report having health tracking apps on their cell phone. In our sample, 39% of the respondents identify as members of the Quantified Self movement.

We asked individuals what kind of data they track using cell phones or websites. Top answers for both include exercise, diet, weight, athletic activity, and sleep (Figure 2). People tend to track more using cell phone apps than they do websites, although both apps and websites were used more than paper or "in your head" tracking.

Our participants tend to self-track more for general health and wellness than to manage a chronic condition. Only 14% of our respondents reported self-tracking primarily for a medical reason. The ranking of types of tracking apps is consistent with this: blood pressure, diabetes, and medication tracking, for example, are much less frequently reported than exercise and diet tracking.

Figure 1. **Characteristics of HDE Individual Survey Participants**

Race: HDE Individuals Survey Sample Compared to US Population

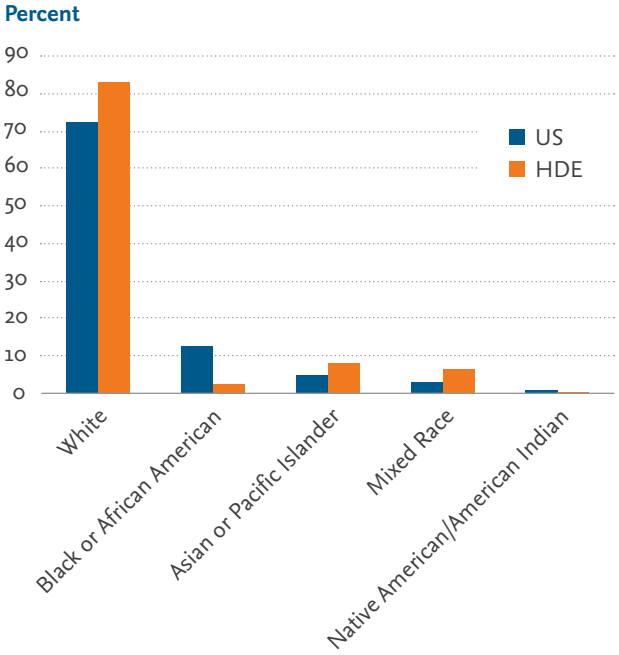
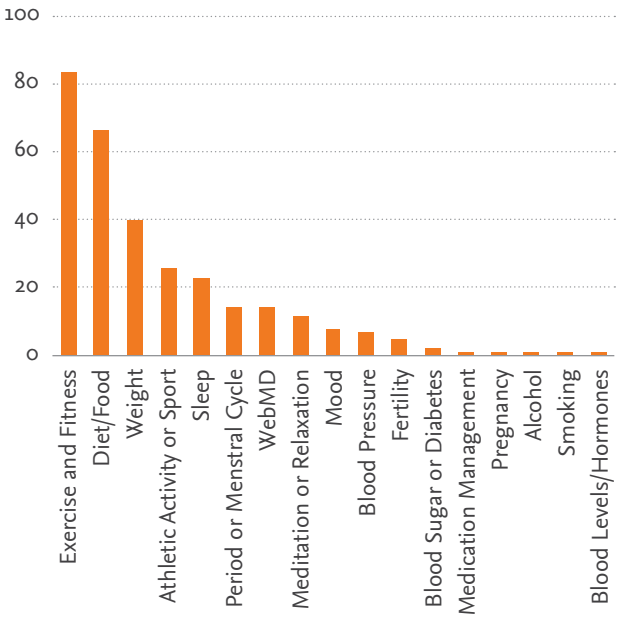


Figure 2. **Types of Health Apps on Phone**

What kind of health apps do you currently have on your phone?



We found the use of cell-phone apps for self-tracking correlated with age, with 100% of 18-25 year olds who track their health using cell phone apps, ranging to only 18% of those age 66 and higher to use cell phone apps (Figure 3). Within our sample, the use of cell phone apps to track health data does not vary significantly by income group.

We asked participants about what they understood with respect to who owns their PHD. We asked the question, "Do you believe that you 'own'—or should own—all of the data that are about you, even when these data are indirectly collected?" Among our respondents, 54% believe they own all their data, 30% believe they share ownership with the company that collected the data, and 4% believe the company owns the data. Interestingly, 13% responded that: "this is not something I care about." We also asked participants, "Do you want to own your data," and 75% said Yes, 5% said No, and 20% indicated they did not care. Ownership is an important concept here because it implies a level of control over the fate of data, and significant portions of our sample both believe they have and want to have that control over their personal health data.

In our sample, 45% of individuals report sharing their health tracking data with someone, either online or offline. Our respondents shared most often with friends and partners, with some of the participants also sharing with health professionals (Figure 4).

Most of our interviewees felt their self-tracking data could be useful to share with their healthcare providers, but that uptake was missing:

> *"I would like to own my data and whenever I go to consult with a professional or a physician or a health care expert I'd like to be able to share that information with them and have them be privy to my entire health record history and I want to monitor it for problems and changes."*

> *"I've talked to my doctors about it and let them know I've been tracking my activity levels. I can see when it's lower than average, or higher than average and sort of try to increase my daily average. They're just like, "OK, that's neat. Sure. You still need to lose weight." I'm like, "Yes, I know!" I feel like to them it's like someone looking up symptoms on Google, and coming up with some crazy illness that they think they have."*

Figure 3. **Use of PHD Apps by Age Group**

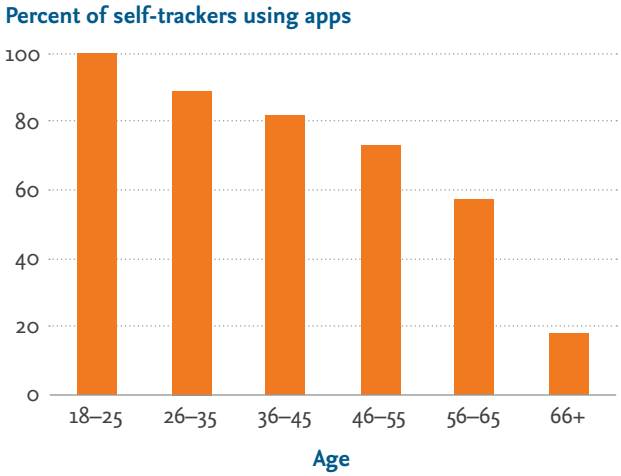
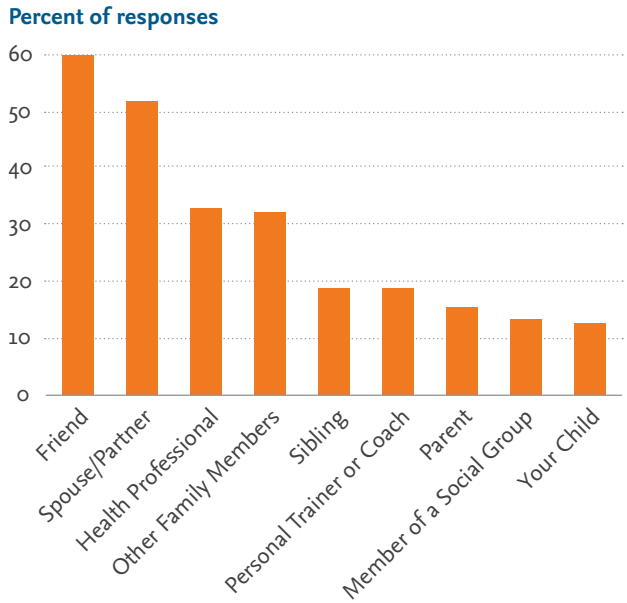


Figure 4. **Who do PHD users share with?**

Who do you share your PHD with?



"My doctor hasn't even requested it. I see him once a year. I'd love for him to, actually, see it. Or, if somehow, even I'm not saying a daily visit, but, maybe even if there was a way that he could look at it, say, for the past...In one snapshot, look at since the last time I've seen him, he could see that I've increased my physical activity."

However, interviewees had concerns about how this might work.

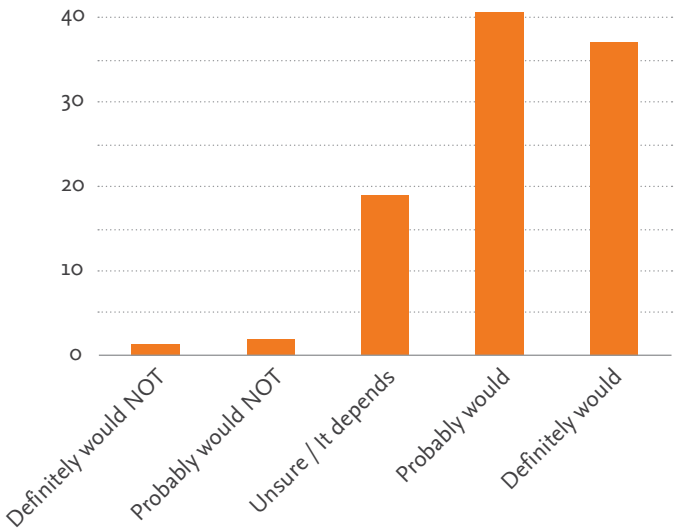
"Once you get into the EMR world, that's got to be completely locked tight, obviously. I have a big problem if that gets used for anything outside of my discussion with my doc."

Our survey respondents reported a general willingness to share their data for use in research, with 78% of respondents answering "Probably Would" or "Definitely Would" when asked if they would be willing to share personal health and activity data with researchers (Figure 5).

Figure 5. **Sharing with Researchers**

Would you be willing to share your PHD with researchers?

Percent of responses



We also asked about willingness to share in two specific cases. Participants in our sample were significantly more willing to share data if it was for a specific scientific study where they had an interest in the topic (χ=14.0, df=4, p=0.007). There was no difference between general willingness to share and willingness to "donate your personal health and activity data to a scientific database." When asked about the importance of compensation, 56% of the participants said that they would be "more" or "much more" likely to share data if they were compensated, and 38% said it would make no difference. We also found that individuals who identify as members of the Quantified Self trend are more willing to share their data for research (χ=24.3, df=4, p<0.001).

For many of our respondents, willingness to share data depends on the purpose for sharing, and many of our participants said they would be more likely to share their data if they knew that it would only be used for public good research. In an open-ended survey question about conditions on sharing, the third most common category of responses (13% of respondents) mentioned an aversion to commercial or profit-making use of their data, with comments including:

"I do not want my data to be shared commercially at all."

"It depends who gets it. Research using these data will be instrumental in the future of personal predictive services, but also for that reason are likely to be exploited by marketers and the politically short-sighted. Thus I would like transparency for who has access to my data."

"NOT NOT EVER for a company to make $$$."

We heard similar sentiments in our interviews:

"Yes, if it was for research purposes, then I'd be interested. If it's for a private agency which is attempting to monetize something about me, then I have no interest."

"If they're using it for research, I don't have a personal problem at all with that. If they're using it for commercial purposes without my knowledge or getting compensated for it, then I have a huge problem with that."

"I guess any kind of corporation or company that would use the information to basically market products, I would feel uncomfortable about that."

On the other hand, some respondents expressed little to no concern about who would use the data:

"I'm not like one of those people who freak out when a company is using their data to increase the value of their company. I get that. That's fine. It's OK. There's probably a privacy policy somewhere that states it that I didn't completely read, and that's totally fine."

Overall, while our participants were cautious about how their self-tracking data would be used, they were generally enthusiastic about the idea of sharing data for research.

"I'm happy to contribute if it could contribute to, say, a larger study where there could be some additional knowledge."

Looking across our data, we find that individuals' willingness to share is dependent on what data is shared, how the data will be used, who will have access to the data and when, what regulations and legal protections are in place, and the level of compensation or benefit (both personal and public).

Our survey and interview results reveal the complexities of the privacy of personal data. First, we found that privacy as a concept is very important to our participants. In our sample, 68% of respondents would only share their data "if privacy were assured," and 67% of respondents said that anonymity is "very" or "extremely" important (Figure 6).

Similarly, in an open-ended question asking participants "Under what agreements and constraints would you share your health and activity tracking data?", 63% of respondents specifically mentioned privacy, anonymity, or confidentiality.

It is worth noting that in each of these cases, approximately one third of the respondents did not see privacy as a major concern. In fact, when specifically asked, 27% of respondents replied that they would share their data without either an assurance of privacy or compensation.

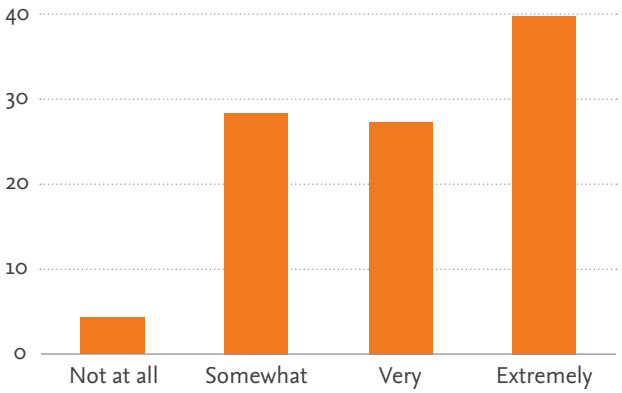
"I would like to own my data and whenever I go to consult with a professional or a physician or a health care expert I'd like to be able to share that information with them and have them be privy to my entire health record history and I want to monitor it for problems and changes."

– Individual

Figure 6. **Attitudes towards anonymity of Personal Health Data**

How important is it to you that your PHD be kept anonymous?

Percent of responses



Our open-ended survey questions and interview data also support this mixed view of privacy. For some participants, privacy wasn't a concern:

> *"It's not really a concern of mine. I mean, to me, it's nothing that's really detrimental to my privacy."*

However, for other participants, keeping their information private is of paramount importance:

> *"Privacy and anonymity is the primary concern."*

> *"So long as you scrub the data for identity markers I would be open to sharing it with any research project that is publicly available."*

> *"I am concerned about privacy and who has access to my information…. The fact that [the app] doesn't store my information online was one of the reasons why I purchased it."*

We also see that participants do not view all data as equally sensitive:

> *"The one thing that might be creepy is if they have like a GPS capability and they could actually track where I'm walking, but to me it's harmless knowing how many steps I've walked."*

Even when our participants believe that privacy is important, they also believe that data privacy may no longer be possible given the pervasiveness of tracking technologies and digital identities in everyday life.

It is important to note that these concerns about privacy may speak more to individuals' attitudes than actual behavior. In our interviews, for example, some participants were unaware of the ways that their data were currently being used:

> *"I don't know. I didn't read their privacy policy or their sharing thing."*

This points to what has been called the "privacy paradox": even when consumers report significant privacy concerns, they often will readily submit private information to companies. Because of this gap between intention and behavior, it is important to treat survey results about privacy with great care (Smith, Dinev, & Xu, 2011).

Some research suggests that while privacy attitudes are influenced by perceptions of the risks associated with disclosure, privacy behaviors are more influenced by perceptions of trust in the recipient of the private information (Norberg, Horne, & Horne, 2007). This relationship between privacy and trust will be addressed further in later sections of this document.

RESEARCHERS

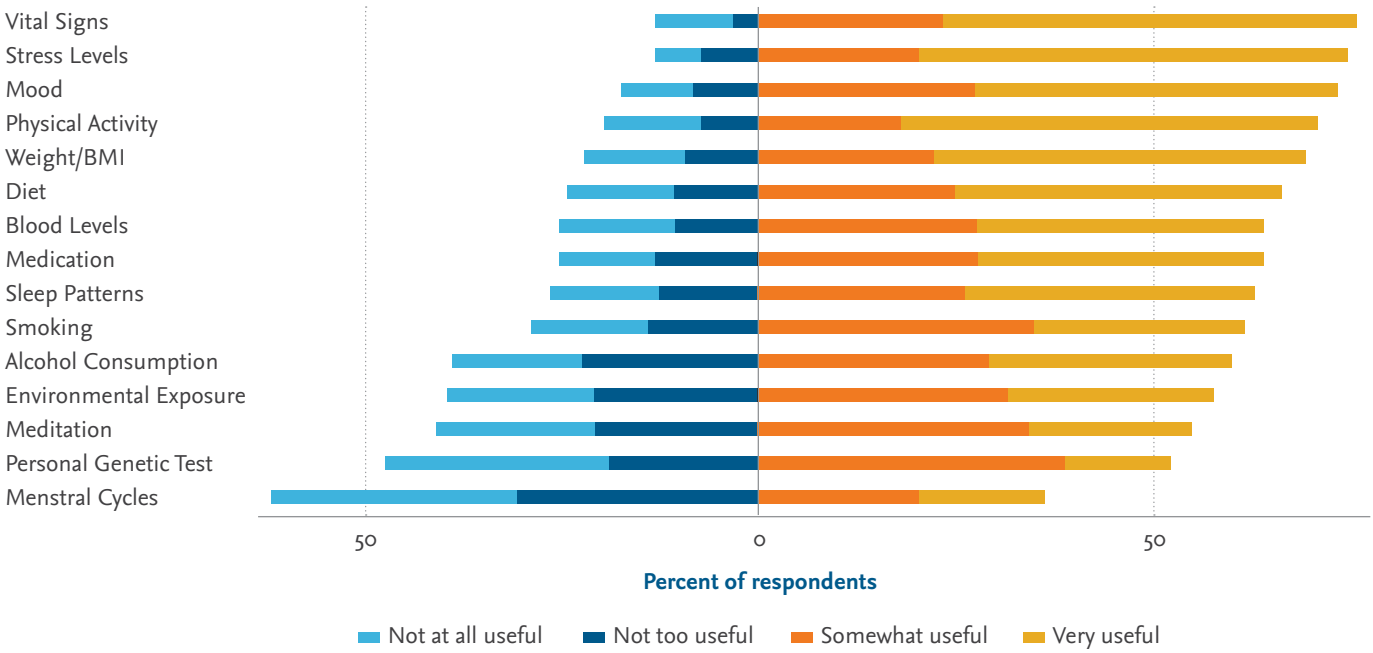
The Researcher survey was taken by 134 participants. Participants come primarily from the health sciences (69%), although social sciences (31%), engineering and technology (19%), life sciences (8%), and arts and humanities (4%) were represented. Respondents were able to select multiple categories to represent their multi-disciplinarity, and 35 did. The most common combinations were health and social sciences (12), health and engineering/technology (8), and health, social, and engineering/technology (5). Seventy-five percent of the respondents were in academia, 11% in non-profits, 8% in government, and 3% in industry. Seventy-four percent lead research programs, 20% conduct research but are not responsible for establishing research goals, and 6% do not currently conduct research. The sample was evenly split among male (49%) and female (51%) participants.

Researchers in our survey were generally enthusiastic about the potential for using self-tracking data in their research, with 89% agreeing or strongly agreeing that self-tracking data will be useful in their own research, and 95% saying that this kind of data could answer questions that other data couldn't. Generally, the categories of data that we found were tracked by individuals will be useful for researchers, although interestingly, some of the most useful research data (vital signs, stress levels, and mood) are much less likely to be self-tracked than activity, weight, and diet (Figure 7).

We also looked at whether researchers in different domains differed in their ratings of data usefulness. The ranking of particular data types does not vary significantly among the health science, social science, and engineering and technology researchers, the life science researchers in our sample (n=8) were notable in that their most useful data categories were (in order): personal genetic test data, blood levels, medication, and diet. In fact, 100% of the life

Figure 7. **Usefulness of PHD to Researchers**

How useful could the following types of self-tracking data be for your research?



science researchers rated genetic data as somewhat useful or very useful, in comparison to 53% of the health science researchers. Because our sample is not representative of a general researcher population, the usefulness rankings should be interpreted with care. However, we are confident in saying that there are researchers who would find each of these categories of self-tracking data to be "Very Useful" in their work.

The potential usefulness of this data was echoed in our interviews, with many researchers detailing the ways that this data can fill in gaps in more traditional clinical data collection.

> *"It doesn't replace what people do in terms of scientific research. I think it just adds another dimension."*

One clear theme was that self-tracking data can provide better measures of everyday behavior and lifestyle.

> *"Right now we're working under a scope of a limited snapshot of people's behaviors that probably isn't accurate. We need to have finer tuned data over longer periods of time to be able to get a better picture".*

One researcher uses self-tracking data to study sleep patterns, and compared self-tracking data to traditional clinical sleep studies:

> *"The thing that's really valuable about this dataset is that there are many nights of sleep, not just one or two. It's in an ecologically naturalistic setting. The person's sleeping at home in their normal bed without all those electrodes. They're getting a more natural night's sleep that's more representative of how they really sleep at home. There isn't the enormous research expense of $1,000 for one night sleep. Having the continuous use of repeated measures makes it possible to investigate not just the variability between people, but also the variability within people."*

Our interviewees also felt that this data could produce research and interventions that were more easily translated into clinical practice and lifestyle or behavior change.

> *"One of the main strengths of this research is that it has potential to be very translational. A lot of the findings that can come out of it can be directly applied in people's lives and are related to the types of health outcomes that people care about a lot."*

It was also clear that for these researchers, aggregating data from multiple sources would be highly beneficial. In particular, linking personal health data with clinical data to provide multiple measures of the same individual was an exciting possibility. One researcher who studies physical rehabilitation outcomes after hospitalization described one possibility for her own research:

> *"The most valuable would be the people who wear the fall devices at home. Just linking that with a simple self-reported questionnaire on health would be fantastic. If you link it both, body weight, even better. If you link it with a full medical record, oh my gosh! We would know so much."*

The survey results also suggest openness to less traditional data sources. Fifty-seven percent have used public data sets, and 19% have purchased data for use in their research. Forty-six percent of the researcher participants have already used self-tracking data in their research, and 23% of the researchers have already collaborated with application, device, or social media companies. Eighty-two percent of the researchers "somewhat disagree" or "strongly disagree" that there are insurmountable barriers to using self-tracking data in their research.

While not insurmountable, researchers did provide examples of the kinds of barriers they face when using personal health data. Researchers found it difficult to negotiate the intellectual property concerns, licensing, and the legal agreements necessary when collaborating with companies. This is a new concern for both the companies and university legal departments. One researcher described waiting months for the university to work out contracts with the company, and in the end, he was unsatisfied with the terms of the agreement:

> *"I think the single thing that would have helped me most would have been if there were some kind of standard data transfer agreement available for this type of purpose that they could have started with as a template instead of whatever they used.... It would have made the whole process take less time, and it would have been a better document in the end, as well."*

Researchers are also concerned about the kind of data that they get from companies. There is very little standardization of sensors, data formats, or practices, making it difficult to understand what the data mean or to aggregate data across multiple sources.

> *"The standardization of the way that data is collected just doesn't exist in a lot of cases. There's too much variability for effective data integration".*

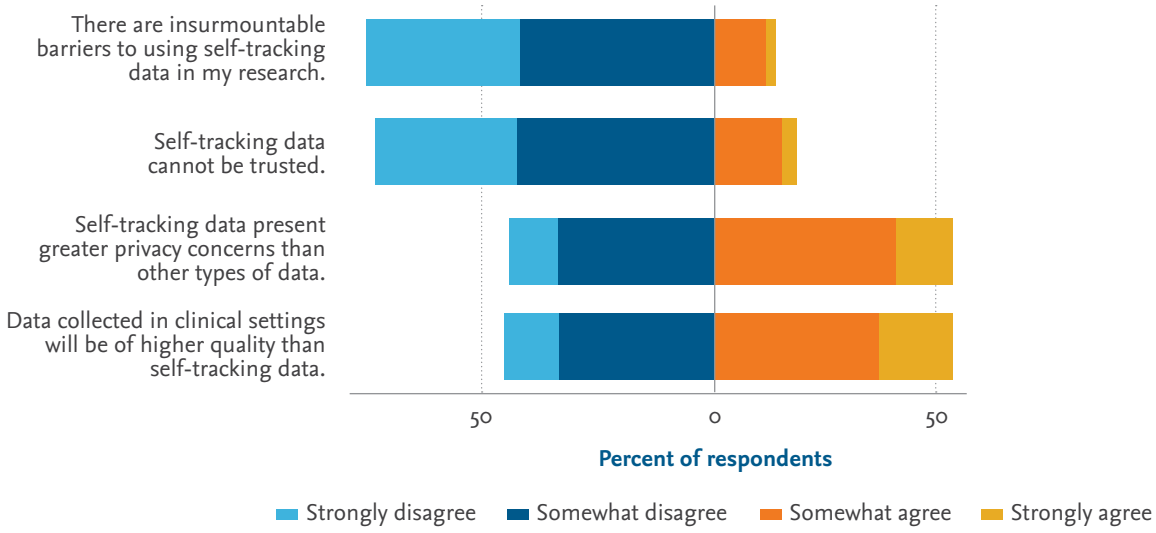
In researchers' experience, companies also tend to be reluctant to share unprocessed data from their devices. In some cases, the algorithms may be proprietary, or there may be other technical reasons that it is difficult to provide "raw" data to researchers. But it is difficult for researchers to understand what they are seeing without low-level device data.

> *"We don't get the raw data that we would like to see. We get data. They've already made a lot of decisions."*

We also asked participants about their perceptions of self-tracking data as research data (Figure 8). Seventy-four percent of researchers somewhat or strongly disagree with the statement that "Self-tracking data cannot be trusted." Researchers who have already used self-tracking data in their research are significantly less likely to mistrust self-tracking data (χ=13.0, df=3, p=0.005), although the only researcher to "strongly agree" that this data cannot be trusted had used self-tracking data in the past. We also found that researchers were split relatively evenly on whether self-tracking data present great privacy and quality concerns than other types of data, and these did not vary with prior self-tracking data use.

Researchers' trust in personal health data stemmed from two sources. First, there is a concern that most of the consumer-level sensors have not gone through any kind of validation process (although one researcher we interviewed was currently conducting a validation study

Figure 8. **Quality of Self-Tracking Data and Barriers to Use**



on one of the market-leading activity tracking devices). Several of our respondents felt that rigorous validation was a necessary prerequisite to using the devices in both research and clinical practice.

> *"There are a few studies of some commercial sensors. Not enough, frankly, to look at their validity and reliability."*

Second, researchers expressed concern about the potential biases in datasets of personal health data. They worried that the users of these devices tend to be self-selected early adopters who can afford the technology, or may otherwise not be representative of a general population. However, one researcher pointed out that no dataset is bias-free:

> *"At the same time, you could make the argument that the people who've agreed to be in clinical trials of any kind are not normal people. I'm not sure it's necessarily much less generalizable than most of the research that gets published. We simply don't have a way of truly representing the population in most research."*

Finally, like the individuals in our study, researchers are concerned about the privacy of personal health data and respecting the rights of those who provide it. However, for most of our researchers, this came down to a straightforward question of whether there is informed consent and how their institutions would handle it.

> *"To me, the whole thing about who should have access to what kind of data really has to do with the person being aware of it."*

> *"I think IRBs are probably really starting to struggle with some of this data collection."*

Overall, the researchers in our study were excited about the possibility of using personal health data. It was seen as complementary to traditional clinical data, allowing them to ask new questions and answer them in new ways. While there are obstacles to using personal health data for research, these do not seem to present greater challenges than any other data source.

COMPANIES AND KEY INFORMANTS

Our interviews with Key Informants revealed concerns in many of the same thematic areas that emerged from our investigations with Individuals and Researchers. For companies operating in this space, advancing research is a worthy goal but not a primary concern. As such, any collaboration with researchers or sharing of research data needs to respect the company's business model and goals. A number of device manufacturers we spoke with, and learned of, view themselves as consumer electronic companies whose primary business is to sell wearable sensors. The data generated from these devices is an asset to help engage the consumer by providing them with meaningful insight. For some companies, especially those that consider the data they collect as a key strategic asset, it is important to keep data out of the hands of their competitors.

> *"Our concern is we don't want our information to end up in the public domain since it's our core intellectual property."*

Another respondent described how a potential partnership with a researcher had fallen apart when the researcher and the company could not come to an agreement about who would own the rights to the results of the research. However, we should note that friction over intellectual property exists across many types of data and is an acknowledged complexity in creating academic and corporate partnerships.

A second dominant category of companies in the PHD space are those that are creating applications that either repurpose data generated by a device manufacturer or allow users to self-enter data. Interestingly, in our key informant interviews, many of these companies, in particular the start-ups, did not view themselves as being "data companies." Even those who were creating mobile applications being used in small traditional clinical trials had little awareness of the potential value of their data to other clinical or academic researchers. One company who described themselves as a health company noted the potential value to them of engagement with this community and noted,

> *"If anything, having research institute academically published on some of the data would help give us more credibility in the market. From a company we are interested in it."*

Companies interviewed also noted that one of the reasons researchers are working closely with industry is the speed at which private companies can make decisions to fund research. Unlike the academic cycle of creating a proposal in response to a solicitation from a federal agency and then waiting six months to hear back on if the proposal was selected for funding, many companies, pharmaceutical in particular, make decisions in weeks. In addition, even when a company is open to donating data to an academic research team the uniqueness of transferring data may cause untenable delay. One company shared with us that it took over six months to get a private research intensive university to approve a standard data sharing agreement where there were no concerns over intellectual property.

The cost to the companies or application developers to share data should also not be underestimated. Many application and device manufacturers have positioned themselves as consumer electronics as opposed to data services companies. The HDE project discovered a great breadth of technical infrastructure and capabilities across the companies interviewed. Even those with technically advanced capabilities may decide not to devote the resources necessary to support data export unless it serves a direct business utility.

> *"Getting data out of our database is not a simple project. The project (with researchers) was going to require engineering resources on our side for something that was not strategic."*

Companies are also very concerned with their relationship with their customers, and sharing data outside of the company presents a risk of loss of customer trust.

At the same time, we also see companies and organizations experimenting with many new models for using these new forms of data for the public good. In some cases, this involves adapting traditional models of sharing data for single studies, with specific and contextual safeguards and agreements. At the other end of the spectrum we see organizations (typically not for-profit companies) working toward completely open datasets using CCo (http://creativecommons.org/about/cco) licenses or fully de-identified datasets. Interestingly, we also found companies that were willing to consider turning over their database to others to run as it grew

beyond their size to support or if the failed in the marketplace. There was also support for the concept of creating a data commons for self-tracking information among a number of companies. Interestingly, a common theme among companies based on their experience of engaging users was that if data donation is going to become sustainable it will need to provide insight back to the donator.

> *"I think we are a small piece of the puzzle and can learn from others."*

We believe it is too early in these experiments to make strong claims about what will be successful, but we are encouraged by the current willingness to try new ideas and models.

One key informant, an academic researcher who also has a strong research relationship with a major company in the PHD space, expressed the concern that unless there was some external source of influence on company practices about sharing PHD for the public good, the focus of the lion's share of corporate research would only be for commercial purposes. This observation aligns with the comments of others that if the field of PHD research is to advance, and if it is to do so based upon the fullest extent possible of data types, a new culture of research will need to emerge that produces win-win situations for all parties.

Another important insight that emerged from the key informant interviews with companies was the importance of user engagement. A number of companies suggested that for data sharing to be sustainable users would need to feel involved, be part of a cause, or gain personal insight from their participation. These mechanisms can create "sticky" practices that engage individuals over the long haul, something needed by both companies and researchers.

Finally, as with individuals and researchers, appropriate use of data that respects individual rights is a key concern. A major finding from our key informant interviews was the importance that trust played in the relationships with their clients.

> *"In terms of user perspective, how you message is more important than terms and conditions. If users are surprised by what you do, you have a problem regardless of what your terms say."*

Companies work hard to build and maintain trusting relationships with their customers, and are sensitive to anything that might harm that relationship. However, this also suggests that when trusted companies decide to participate in data sharing with researchers, it could be seen as a powerful endorsement by their user base.

While there was no consensus on the best approach, our key informants, more than our other cohorts, highlighted the complexity of privacy, informed consent, and personal data. What became clear was the deep intertwining of data privacy, IRBs, informed consent, licensing agreements, network and database security, HIPAA and other legal frameworks (both national and international), user interface design, corporate policies and customer relations.

4.4 Vignettes

The following vignettes are offered as a complement to the survey and interview findings and convey our overall understanding of how individuals and researchers consider the area of personal health data. These vignettes are composites developed from our interviews with individuals and researchers, and from open-ended survey responses.

Individuals

Rhonda is a busy professional with an advanced degree. She mostly tracks her activity with a paper planner, blocking out physical activities like yoga, hiking, or aikido in her weekly schedule. She can make a quick read, visually, about her relative levels of activity each week, and uses this information to make sure she plans a hike or vigorous exercise for the weekend if she doesn't think she or her partner have been active enough. He doesn't self-track, and Rhonda uses a Jawbone UP mainly to provide moral support for a friend who started using one to meet specific activity and weight loss goals. Initially, she used the device's diet tracking function but found that her caloric intake rose and fell in parallel to the calories she burned each day. She didn't find it worthwhile to continue with that tracking. She also discovered that the apparently innocuous data could be revealing. Her friends, with whom she shared her data profile, could infer moments of intimacy from her sleep cycles. She would happily share her data for research for the public good. Like some of the interviewees, Rhonda feels that privacy is a thing of the past. She feels that sharing anonymized, aggregated data is pretty risk free, but believes that university research requirements and peer review would protect her data from being used unethically.

Arturo is a young professional working in the non-profit sector. He is an avid self-tracker, and leads an active lifestyle. He uses a Fitbit and multiple apps. Arturo is curious about how these apps work to incentivize behaviors, and experiments eagerly with a variety of apps. One app that he uses allows him to compare his mountain bike rides to other users with similar physical statistics riding on the same course. He compares his Fitbit data with people he knows, and has a friendly competition with his father. Arturo knows that this kind of data could be really useful to researchers, and already shares his Fitbit data with a national health study. Arturo would share is data widely: "As far as I'm concerned, the whole world can have it as long as it's anonymous." Like other interviewees, Arturo thinks the concept of privacy might be moribund, a trade-off for other benefits. He shrugs off his concerns, saying "If the data was used to sort of pinpoint me as a specific demographic user I wouldn't like it, but I also see it as an inevitability. I feel like it's pointless to argue against it, because it's a runaway train. I don't see how it can be stopped. I continue to use Facebook." Partly, his laid back attitude about privacy stems from his beliefs about what the data say about him: "You know, I could spend a lot of time worrying about my data privacy on this kind of stuff. But if a life insurance company was going to look at it, they'd look at it more favorably. But for someone else it could be a very big deal to have that stuff out there. I personally I don't have that view, but I can absolutely understand why someone would."

George works in software development. Like Arturo, he uses multiple devices and apps to track his personal data. A young baby boomer, he has some chronic health issues he manages, in part with these devices. Aware of the limitations and contradictions of tracking, George says he "likes monitoring these things. I do this. It's just the tip of the iceberg. These are just monitored by me because they're the only things I have a handle on. It's like a drunkard looking for his keys under the light." Still, he has high hopes for how these technologies could change his relationship with his physician as well as make a difference for health care reform. "These are the only tools that I have. But quite frankly I'd like to have all of my health and medical records of any kind, including imaging data and test results, everything, under my own control." As it stands, his doctor isn't interested in seeing the data he collects, a frustrating situation that several interviewees shared. He'd like to have ways to share his data automatically to facilitate his health care, but knows it's not a simple matter: **"**Once you get into the electronic medical record world, that's got to be completely locked tight, obviously. I have a big problem if that gets used for anything outside of my discussion with my doc." George knows his data is valuable, and wants to share it with researchers working in the public interest, but with restrictions. He hopes that this kind of research will lead to new infrastructures for sharing with healthcare professionals and to make real-time adjustments in his self-care.

These vignettes show the complexity of the space of personal health data. For some, individuals, self-tracking is a tool to live a healthier life, but we also see how these same practices and technologies can be used to monitor chronic medical conditions. While we refer to this as *personal* health data, these vignettes also reveal that the data is deeply social. Self-tracking can not only help individuals understand themselves, they can be important relational tools, supporting and enriching friendships, providing a venue for friendly competition with a family member, or, potentially, helping to create common ground with health professionals.

RESEARCHERS

Lois is a university medical school-based researcher working studying cardiovascular disease. "Gold standard" data in her field requires expensive laboratory tests that only provide data from one or a few time points for each individual in the study. In order to fill in gaps in what she can see from clinical data, she worked to cultivate a relationship with a company whose device collects heart rate data. She is eager to continue doing research with datasets like this, but has found herself trailblazing paths at every step. Getting the data from the company proved to be a challenge. The company needed evidence of her IRB approval, which was relatively straightforward because the data was already deidentified. However, hammering out a data transfer agreement between the company and her university became a headache. "The lawyers at my university had to negotiate with a lawyer at the company. That was a slow process, and I didn't actually have much say in what the agreement ultimately contained." As a result, the agreement does not address many of Lois' concerns. It can take years to move through the process of analyzing the dataset, writing and submitting publications, and shepherding them through peer review. However, the data transfer agreement allows the company to terminate the agreement at any time, and Lois is worried that she might lose access to the data at a moment's notice. She also worries that this kind of data is so new that she might face resistance from peer reviewers: "Whether or not it will be published is a whole other issue." Even so, Lois is excited about using this data. "There's potential to discover a better understanding of how lifestyle affects health. Lots of people are trying right now to manage all sorts of different symptoms through lifestyle, but a lot of the information they have is basically hearsay on the Internet. This kind of research is more and more about the things that really work, and putting the information that people need into their hands."

Stefani is an assistant professor in public health, leading NIH- and NSF-funded projects studying the efficacy of lifestyle interventions for treating obesity, diabetes, and other chronic health conditions. Stefani has been using self-tracking data in her work for years, but from devices intended for medical use. "I'll use them for a few days at a time up to a week or two weeks across an intervention period." She is excited about the potential of using consumer-level tracking in her research. "To me, the goal is long-term data collection of multiple health behaviors. Right now we're working under a scope of a limited snapshot of people's behaviors that probably isn't accurate." But Stefani worries about the quality of the data that come from consumer-level devices. "I'm interested in accurately measuring behavior, so I would tend to use more of a research grade device with greater validation." The lower cost of consumer devices makes them available for wider use, and their connections to smartphones or web-based software can provide a platform for interventions. However, until there are validation studies of the devices—both that they are collecting good data, and that people tend to use them as expected—she is not sure she can trust the data and doesn't think they will be accepted by the research community. In the meantime, Stefani is conducting a small validation study of a consumer-level device, and would like to help companies produce better devices. "I would like to partner with a company that has developed or would like to develop a great personal monitor that collects raw data that can be shared publicly." Stefani believes that her experience could help a company produce a better device, and that she could help ensure that it would produce high-quality, transparent data in an ethically responsible way that would make it easier for researchers to use.

Like most of our Researcher participants, both Stefani and Lois are excited about the potential of using personal health data in their research. Lois has faced a number of organizational barriers in order to work with an external company, but the data she received has been extremely useful in helping to understand longer-term behavior. For Stefani, data from consumer-level devices has (so far) been too problematic to use in her own work. On the other hand, she is eager to work with companies, not only to get the data, but also to help them produce better quality devices and lifestyle interventions. For academic researchers, we also see that publication continues to be the metric by which success is measured, and while the riskiness of using a new data source has not stopped these researchers, they still worry about whether the results of their work will be accepted by their wider research communities.

COMPANIES

DeviceCo is a large manufacturer of wearable devices, having sold millions of units. DeviceCo's product is just one of many consumer electronic commodities it manufactures. Since DeviceCo understands itself to be primarily an electronics company, the in-house research team focuses on using the self tracking data collected to improve the device and user experience by generating useful insights for users. Because the product is so popular, researchers are very interested in partnering with the company. On a few occasions, DeviceCo has worked with researchers to share data, but has discovered that working out the details of the partnerships is more complicated and time-intensive than might be expected. At least from DeviceCo's, perspective, researchers were able to benefit from these partnerships and published papers about research with the datasets they shared. But the benefit to the company was not clear, and due to the costs involved in working with researchers, DeviceCo has not been eager to collaborate with researchers. Still, DeviceCo understands that the data has untapped research value is open to partnerships in the future if either costs can be reduced or a benefit realized for sharing, or possibly a mixture of both.

HealthStartup's three co-founders hope to help others suffering from Condition Z by inviting people to share their experiences of the disease and treatment efforts that have helped. The intent is to crowd source sets of "best practices" for treatment and diagnosis based on these experiences. HealthStartup has been more than modestly successful in that endeavor, but the founders noticed that Condition Z users of HealthStartup have also been interested in gathering data to make informed, evidence-based decisions as patients. HealthStartup, like many companies operating in this area, has been approached by researchers who'd like to work collaboratively in this new direction. Developing proposals for working together has turned out to take a lot longer than anticipated, time that is difficult for a startup that must carve out its niche quickly in order to survive and grow. While HealthStartup would like to forge these relationships, it has also been approached by private companies, including some in the pharmaceutical industry, to do similar kinds of projects. Those private companies are equipped to move much more quickly. Not only must HealthStartup think about its bottom line, there is a sense of urgency in its mission to help its clients who are living with Condition Z. The mismatch in time frame between researchers and startups like HealthStartup has meant working in the short term, rather than planning for the long term, on this data research.

4.5 The Personal Health Data Ecosystem, 2013

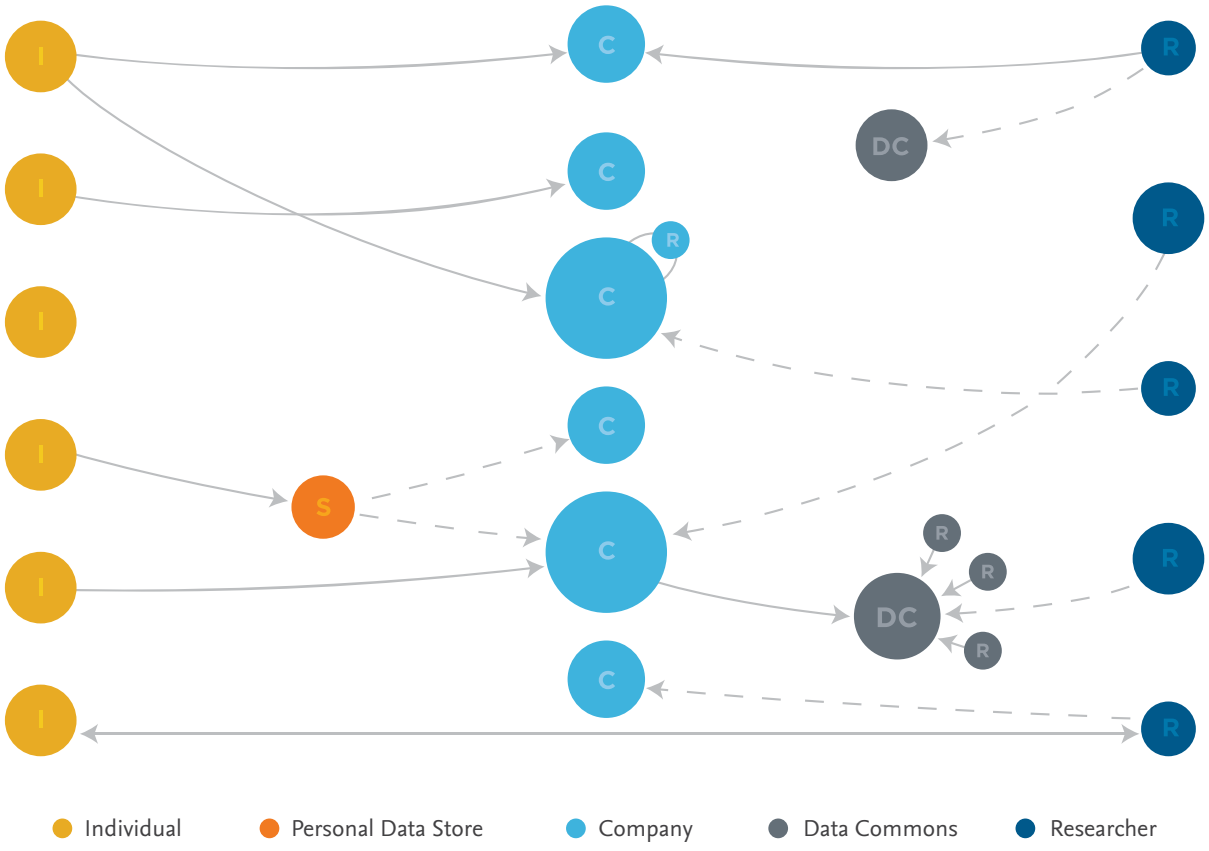
As a result of our survey and key informant interview, we present the following conceptual overview of the many approaches being used to capture and use PHD for research. One of our key findings is the breadth of current activity occurring in this space (Figure 9).

Individuals are currently sharing data with companies who are providing them with devices or applications, while a few early adopters are experimenting with personal data stores or sharing their data directly with researchers in a small set of clinical experiments primarily focused at patient verses population level.

There are a prolific number of companies in this space. Many of the larger companies have their own research staff analyzing user behavior but this tends to be done to improve product services as opposed to create scientific knowledge. There are a small number of companies, primarily those with more of a health research focus, who are trying to develop data commons to regularize data sharing with the public and researchers. However, most companies are not yet regularly sharing data with academic researchers. When these relationships do exist, they tend to be fragile and built on personal relationships. It is not easy for most researchers to gain easy access or attention from companies that have PHD.

Researchers, primarily individuals, not yet full research teams, are beginning to experiment with PHD data but at the level of one off access to unique data sets that are not more broadly accessible to the community. Interactions with companies and the rare data commons tend to once again be more based on personal relationships than any open data sharing frameworks.

Figure 9. **Personal Health Data Exchange and Use for Research**



5. Key Issues for Personal Health Data Research

In the course of this project several issues emerged that are fundamental to research conducted with PHD. While these are common to many types of medical, behavioral, social science and public health research, several unique challenges arise when considering these in this new ecosystem of personal health data generation and inquiry.

5.1 Privacy and Anonymity

Privacy and anonymity emerged as key issues deserving special consideration in the Health Data Exploration project.

Privacy is a complex and critical issue that needs to be addressed to develop the appropriate methods for sharing self-tracking data with the research community. One framework for better understanding privacy involves understanding the "contextual expectation" of the user. Three critical parameters can be examined: the actors (subject, sender, recipient), attributes (types of information), and transmission principles (constraints on the flow of information). Understanding these elements help foster the development of normative behavior for how information should be shared (Nissenbaum, 2011).

This framework can help identify the sources of complexity of privacy in relation to self-tracking data. Consider for just a moment the breadth of information (attributes) covered by self-tracking. Data ranges from personal impression of mood to device-generated measurements of physical activity and scientific clinical measurement of blood and genomic data. Each data type may elicit unique user expectations regarding privacy. However, digital sharing with academics has not occurred long enough for normative behaviors to emerge, and expectations remain heterogeneous. Put simply, we do not yet know the contextual expectations of privacy associated with individuals who self-track.

While we may not yet know enough to understand the full contextual expectations for privacy, we do know that is a key concern among individuals who are willing to share their data with researchers. The HDE survey revealed that about 70% of respondents would be willing to share their data with academic researchers with the dominant condition (57%) for sharing being an assurance of privacy for that data. Importantly, the survey also found a considerable cohort of roughly 30% for who privacy was not a consideration with regards to sharing. The company and key informant interviews show the potential for these data to carry a high level of personal attachment. One large device manufacturer noted that some of its users consider their physical activity data to be more private than a blood test.

Individuals from the HDE survey are also clearly concerned about the anonymity of this data. Over 90% of respondents said that it was important that any health and physical activity data they shared be anonymized. A national survey recently completed by Pew Foundation focused more general on on-line privacy reveals a growing general concern about digital anonymity. Pew found that 86% of survey respondents had taken some steps to either remove or mask their digital online behavior. Interestingly, after the "Summer of Snowden," the dominant concern expressed was not over government tracking but rather access of this data by hackers, advertisers, or friends and family. Pew Study Director Lee Rainie summed this sentiment up by noting, "Users clearly want the option of being anonymous online and increasingly worry that this is not possible." (Rainie, Kiesler, Kang, & Madden, 2013)

Given a requirement of anonymity for sharing data with researchers it is necessary to examine whether this condition is easily achievable. The last five years have seen a growth in academic research that demonstrates the various commercial, mathematical, and linked data methods that can be used to re-identify anonymously

shared data. Sweeney and her colleagues at the Data Privacy Lab at Harvard were able to identify between 84-97% of anonymous profiles in the Personal Genome Project database using metadata including birth, gender, and zip of users (Sweeney, Abu, & Winn, 2013). Database size is also not necessarily a deterrent to re-identification given that many human behaviors create patterns that are highly unique. Recent research analyzing cell-phone data for 1.5 million users showed that with as few as four spatiotemporal points of data researchers could identify 95% of individuals (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013). We live in an era in which advanced computational techniques and data mining approaches are substantially challenging the concept of anonymity. The honest and frank answer to the question of whether anonymity of digital shared data can be guaranteed is no. However, by using emergent best practices (avoiding some types of metadata (zip code as an example) and scanning uploaded files for "name fragments") we can make re-identification more challenging.

Privacy issues are complex and emergent in relation to self-tracking. However, enlightened conversations about privacy, anonymity, and the contextual expectations related to self-tracking data are an essential foundation for building an ethical ecosystem that encourages individuals to donate their personal data while respecting their rights. Some proponents have noted that the questions "for who, when, and for what purpose" are part of an essential social justice conversation that balances personal rights with competing uses for this information (Neff, 2013). Left to market forces alone an imbalanced ecosystem could occur, resulting in unfettered mining of personal data and creating public backlash (World Economic Forum, 2011).

Based on the research conducted for the HDE project, we believe that these critical issues need to be addressed by a multi-stakeholder community that involve individuals who self-track, companies creating devices and storing data and academic researchers. First, additional research is needed to help unpack and understand user expectations regarding the privacy of self-tracking data. This understanding can then help inform conversations regarding establishing norms of use. Second, there is a need to develop appropriate education and outreach materials help discuss the realities and challenges of digital anonymity. Third, tools need to be developed to enhance users' control of their data, awareness of sharing, and notification of findings derived from data use. These controls are an essential condition for establishing the trust needed to assure that data donation is not a one-time occurrence.

It is unclear the extent to which existing laws provide privacy protection to self-tracking and PHD. There is no direct right to privacy in federal law. Rather, in the US there is a patchwork of laws governing privacy for specific types of data (patient billing, vehicle registration, education records, video rental) (Singer, 2013). In the medical context, the Health Insurance Portability and Accountability Act (HIPPA) created a new privacy right for personal health information (demographic information, medical history, test results and insurance information in the medical record) within the medical record (Rouse, 2010). Requirements for handling this data found also only covers regulated entities (health care providers, health plans, health care clearinghouses (45 CFR 160.102, 164.500). As such, sharing of data by a patient of their own health information, even including medical tests, voluntarily in a social network like Facebook is not covered. In addition, even standard Fourth Amendment protection (against unreasonable search and seizure) is voided if the private information is shared with a third party (Asprey, 2013). Said differently, all data willingly shared with a device manufacturer (Nike, Fitbit, BodyMedia, etc.) has no Fourth Amendment coverage.

One of the dominant concerns expressed about donating data is the risk to the individual if that data is identified. Not dissimilar to the quilt of privacy laws, protection from personal being used against the individual is based on data type and use. For example, the Genetic Information Nondiscrimination Act (GINA) protects against health insurance and employment discrimination related to genomic information. However, the law does not apply to changes that could be made to your life, disability, or long-term care insurance based on DNA information (National Human Genome Research Institute (NHGRI), 2010).

5.2 Human Subjects Research and Informed consent

In response to historical ethical failures involving human subjects, an independent review process for human subjects experiments was created in the US. Federal law mandates the creation of Institutional Review Boards (IRB) for review of research proposals involving human subjects and using federal funding (CFR 45.46). The Belmont Report, finalized in 1978, clarified fundamental ethical considerations for IRBs when reviewing human subject experiments. These three categories of concern include: 1) Respects for Persons, 2) Benefice (no harm to the individual, maximize benefit) and 3) Justice (balance of risk and benefit) (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (NCPHS), 1979).

The current academic practice is to use IRB review for all human subject experiments regardless of funding source. IRBs have a minimum of five representatives from different academic disciplines with efforts taken to balance gender and a requirement that one of the members be external (45 CFR 46.107). Each research university usually has its own IRB, and within the last decade, some institutions have created distinct IRBs for life and biological sciences and social sciences. IRB members determine if the proposed research is in compliance of with the law and ethical guidelines and may approve, reject, or request modification to all research proposals.

Is self-tracking research likely to be thought of as human subject research by an IRB? All of the key informants we interviewed indicated that their self-tracking research projects have undergone IRB review. The degree of review has varied considerably depending on the research project. IRB outcomes varied including waivers of IRB review (determined not to meet criteria for requiring human subjects review), granting of expedited review (done by single IRB member, determined to be of minimal risk to individual), and full review (requiring documentation of informed consent by study participants).

Certain types of self-tracking data clearly require full IRB review. For example, most researchers would agree that research drawn from personal medical records requires full review. However, there would likely be substantial disagreement among researchers about the IRB approval needed to conduct research using fitness activity data posted by Fitbit users. The second example would include assessing if the data was already public, given that it was shared by the subject with the company, determining if downloading the data constitutes an interaction between researcher and the subject and evaluating what risks exist if that personal data was disclosed.

Some of ethical issues related to self-tracking academic research have already been explored in the area of Internet research. The rise of the Web, blogs, social networks, and massively multiplayer online games ignited academic research that raised issues about the existing paradigm of evaluation used for human interactions traditionally used by IRBs. IRBs seem about equally split on the question of whether Internet research raises unique ethical concerns with 50.3% of institutions agreeing they do and 47.6% saying they do not. Yet, most institutions (~72%) have no formal guidelines for research dealing with this type of data (Buchanan, 2010).

While there are not yet standard guidelines for using Internet data, researchers in this area have made important contributions that help frame emerging issues. Two contributions from the area of Internet research ethics to self-tracking are the concepts of human non-subjects data and the human harming research. Human non-subject data is a new categorization proposed for de-identified human data. Proponents suggest that this category would not necessarily need full IRB review and could instead use a set of best practices to minimize re-identification and give subjects the ability to opt-out of research projects (Brothers & Clayton, 2010). The second concept, human harming research, has to do with a shift away from traditional methods of assessing risk to subjects. Traditionally IRBs have used a metric for assessing harm based on the direct interaction between researcher and subject. Some computer security researchers have argued that the proximity test used by IRBs needs to change to reflect the realities of the digital age. They propose that the ethical assessment should be reframed to focus on the potential for the research to harm humans. This shift in paradigm would help raise awareness of ethical

considerations among a cohort of academic researchers (those in computer science) who traditionally have not had considered human subject issues and, over time, create more useful conversations regarding risk by IRBs (Buchanan & Zimmer, 2012).

Human subjects research requires the "informed consent" of the proposed subjects. This requirement is based on a primary ethical consideration of autonomy of the individual and the rights of individuals to determine what will happen to them. The Belmont Report developed three elements for use in the informed consent process, including the need to share detailed information about the project with the subject, for that subject to comprehend the nature of the experiment and any risks, and for the agreement to be entered into on a voluntary basis (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (NCPHS), 1979). The Department of Health and Human Services has codified these principles into an informed consent checklist for use by researchers that elaborate on the considerations, documentation, and conditions under which a waiver may be granted. Waiver conditions include: (1) research that involves minimal risk to subject, (2) the waiver does not affect the rights or welfare of subjects, (3) it is not practicable to carry out the research without the waiver, or (4) subjects receive pertinent information after the study (45 CFR 46.116).

The HDE key informant interview and related experience of the research team reveal a number of different approaches for informed consent involving self-tracking research. These approaches, somewhat similar to the judgment by the IRB of whether the project is human subject or not, cover the full spectrum from documented consent to full waivers of the requirement. Waivers were often granted due to an assessment of minimal risk to the individual or a judgment that it was impractical to get consent from a large population.

It will be useful for the PHD research community to examine how other disciplines have dealt with the issue. Both clinical and genomic researchers have struggled to apply traditional models of obtaining consent for their large human data sets. Tension has existed in the need to respect the autonomy of the donor while trying to promote the maximum scientific benefit from a data set. A key challenge for this group has been the re-use of data. It is necessary to note that it is not possible to achieve informed consent by a subject to all future uses of their sample (Arnason, 2004). This is in large part due to an inability to inform the subject about what all of those future uses might be.

MODELS FOR CONSENT

Several models for understanding and obtaining consent have emerged in response to new kinds of research over the last 15 years. One new framework developed to address the challenge of informed consent is open consent. Open consent (OC) requires that volunteers who donate personal genomic and health record data do so with an understanding of risks to themselves and without any guarantee of anonymity, privacy, or confidentiality. The Personal Genome Project at Harvard has pioneered this concept (PGP, 2013). Participants who agree to donate data undergo extensive on-line testing to demonstrate their understanding of the OC agreement prior to sharing their data. The model is based on the argument that transparency of purpose and veracity (truth telling to the subject regarding risks) creates a process that is as "fully informed as possible." (Lunshof, Chadwick, Vorhaus, & Church, 2008).

The Consent to Research movement has used open consent as the basis for creating what is called "portable legal consent." This approach creates a lengthy and thoughtful process for subjects to volunteer their data for research, including reading lengthy documentation, viewing on-line tutorials, and signing a document. The portable legal consent document details the study purpose, procedures, risks and discomforts, benefits, confidentiality and a grant of permission to use gathered data until 2080 unless the user decides to depart from the study in writing. The Western Institutional Review Board (WIRB), an independent IRB, has approved this form of informed consent ("Consent to Research," 2013).

Another model for obtaining consent, created in Europe, is discrete consent. This model rejects the notion of broadly donating data and instead focuses on individual involvement in approving each potential use. Discrete consent involves an interactive and dynamic infrastructure that notifies individuals of each potential use of their data and then empowers

them to choose to share or not share. The model rejects the current status quo of "one and done" for sharing arguing that these systems give individuals no real control of their personal information. The Ensuring Consent and Revocation Project (EnCoRe, http://www.encore-project.info), developed with support from Hewlett-Packard in the United Kingdom, created a technical infrastructure to support this vision, including software assistants to allow subjects to express their privacy preferences, and a centralized repository of data with policy, audit, and trust authorities (Mont, Sharma, Pearson, Saeed, & Filz, 2011).

Given the range of data and research activities in self-tracking it is infeasible to frame a general answer to the applicability of IRB approval or informed consent to self-tracking research. However, as research grows in this area there will be increasing friction in the continued application of pre-digital concepts for dealing with human subjects. The HDE survey and key informant interviews with individuals, companies, and academic researchers have identified trust as an essential element in data sharing. Given this critical role, we believe it is important, even if not legally required, for proposed self-tracking research to undergo IRB review. We do not preclude that this assessment may result in a determination that the research does not involve human subjects within the operational definition of IRBs. While this framework is not without it flaws, it is the single best existing framework to protect the rights of the individual against unethical experimentation. These safeguards are necessary to avoid damaging the implicit trust that exists between the public and the academic research community essential to sustain the donation of personal data for the public good.

5.3. Data Sharing and Access

In order to understand the landscape of current data collection and sharing practices, a review of several websites and tracking applications was conducted. We sampled from popular websites with millions of users, sites already generating health research from personal data, device manufacturers, and entities with innovative sharing models. Research consisted of a detailed review of Terms of Service and Privacy Policy documents, interviews with key informants and secondary sources. Several of these are analyzed as exemplars.

INNOVATIVE MODELS

The data sharing models of three websites are described below. These websites share several common characteristics. Most are focused on sharing data between individuals and researchers. They tend to be transparent in their terms and business model in this regard. They have built in user protections, such as opt-ins or informed consent. Many provide users with a relatively high level of control over their data.

Personal Genome Project is an open, not-for-profit online repository of genetic and other health related data. It originated out of a research project at Harvard and is specifically focused on providing a public repository without commercial motivations. PGP is notable for its extensive consent process, which presents detailed information about the uses and risks of posting such data and requires completion of an enrollment exam to promote understanding. Data can be submitted in a wide variety of formats and is openly available for public download. It is one of the few sites to specifically assert that it does not own the data and instead makes the data available using the Creative Commons CCo 1.0 Universal waiver. While names are not associated with publically available data, PGP warns users that third parties may nonetheless be able to identify individuals.

PatientsLikeMe is a private company that collects information related to chronic disease. The focus is on users submitting self-reported metrics to support research efforts. While personally identifiable information is restricted, the intent is to share all other submitted data. The Terms of Service describes data recipients such as "pharmaceutical companies, medical device companies, non-profits, and research institutions". Data from the site has been used in over 20 peer-reviewed scientific articles and there are several opt-in options such as allowing data recipients to directly contact users and restricting visibility to registered users.

23andMe is a commercial website that allows consumers to submit personal samples for genetic testing. The website allows consumers to augment submitted genetic data with self-reported data such as disease traits or demographics. Users can opt in at different levels of participation by choosing to submit anonymized data at the aggregate level or individual-level data if they choose to. The terms of service is explicit in describing that qualified researchers are potential recipients of this data.

Data may be transferred to those partners or accesses on-site at 23andMe under more restrictive circumstances. (Note: As of the date of this report 23andMe has been prohibited by the FDA from marketing its service in ways that imply that it provides medical advice. The FDA has requested further clarification about how 23andMe cautions users against over-interpretation of the results.)

TERMS AND POLICIES

A review of a larger set of policies, in addition to examining the specific cases above, revealed several dimensions that are relevant to users who share health data. While the language of these policies is informative, it is important to bear in mind how these policies may differ from reality. In interviews, key informants stated that they would be cautious of any behavior that might erode user trust or satisfaction, even if such a behavior was explicitly allowed according to their terms of service.

Rights, ownership and licensing are all terms that relate to what control the user and receiving entity have over data. Other than OpenPaths and PersonalGenome-Project, none of the reviewed websites or applications make use of the term "data ownership" in their Terms of Service or Privacy Policy. The most common element is a complete, sub-licensable, irrevocable license of "User-generated content" to the receiving party. User-generated content typically refers to content such as posts, messages and photos. For some websites collecting self-reported data such as weight or exercise, it is not explicitly stated if this information falls under "user-generated content" and that content's license. At least one activity device manufacturer stated that it has rights to all content that is "derivative" of its services, which may apply to the activity data itself.

Most but not all of the reviewed policies are reasonably detailed with respect to what data are being collected and with whom they may be shared. Common categories of data collection include demographics, weight and other body metrics, and survey responses. Potentially sensitive categories include genetic data, family history, contacts and social networks and GPS location. Almost all policies describe the need to share data with third parties in order to fulfill business operations (e.g. payments, customer service). Most also include advertising or marketing partners and a handful specifically mention pharmaceutical and medical device companies. Almost half of the policies specifically mention "researchers" as potential partners. In some cases this is as brief as saying that aggregate information may be shared, while the websites reviewed above tended to be much more explicit in describing the scope and process for that research. Lastly, three policies specifically stated that user data could be sold to other parties.

Deletion of user data varies widely among policies. Typically, a deletion request must be manually submitted through email or customer service, as opposed to an automated process initiated online (e.g. that provided by Google Accounts). Less than half of the policies address the ability to delete, and several of these warn that personal data will likely remain in archive form. One device manufacturer stated that following a request for deletion, the data might still be retained and used in an anonymized form. Two of the more research-focused sites reminded users that data could not be deleted from completed studies.

APIs

Websites and apps are increasingly offering technical interfaces for downloading, querying and possibly modifying data. APIs (application programming interfaces) are the specifications for the commands to perform these actions. APIs allow third party developers to build new applications that interact with the exposed data. They also allow tech-savvy users to interact directly with their own data. Of the 19 companies reviewed, 12 mention some form of API.

The accessibility of APIs can vary and is a critical consideration in the data sharing model of a website. An API that is described as relatively open might include clear documentation, robust access to data and an open registration process for becoming a partner. Companies fitting this profile include 23andMe, OpenPaths, and Fitbit. Successful APIs can lead to a large number of registered partners. Withings has 80 plus partners and HealthGraph (the engine behind RunKeeper) has approximately 120 partners. These developers argue that the costs associated with offering an API are outweighed by the benefits, such as added functionality provided by third parties and increased enthusiasm in the user base.

However, other companies provide relatively limited APIs. Several companies have closed registration, meaning third parties must be invited or go through a selective application process. Additionally, APIs may provide lim-

ited functionality. The API may allow users to add to but not extract information from their profile. Alternatively, an API may provide access to high-level information like total steps taken for an activity tracking device, but not the raw accelerometry data.

In summary, APIs are an important aspect of the sharing model of a website or application. A strong, open API can provide the control that is typically associated with data ownership. An API provides an efficient way of connecting and transferring data, whether it concerns users linking their individual tools or researchers aggregating large datasets. That being said, APIs are just one part of the sharing model for sharing. Website can still provide users with robust control over data through the standard interface, customer service and effective policies.

PERSONAL DATA STORES

The plethora of devices and applications used by individuals for self-tracking create issues related to data consolidation and control. Within the last eighteen months, a number of new companies began providing users with the tools they need to create their own unified data dashboards. One newly emerging company in this niche is Human API. Human API empowers users to aggregate their own data from up to 50 manufacturers while storing the results in a HIPAA-grade security private cloud.

One emergent architecture for addressing the challenges of data control relevant to PHD is personal data stores (PDS). A PDS is a user-controlled datastore that has the ability to seamlessly share data with third party applications through an API that controls permissions (Windley, 2010). This architecture signals a fundamental shift in which the user becomes the point of data integration. The user is in control of what, when, and with whom data shared. Not surprisingly, PHD has been a major use case behind the development of these tools.

While the PDS concept is still early in development and implementation, there are still a few initial software development projects worthy of note. The Locker Project (http://www.lockerproject.org) is an open source software project that allows users to aggregate a great deal of their personal information from various sources into one user-controlled database. The primary software developer for the project then co-founded the startup Singly as a company to help mature the code and push forward with application development in health, digital photos, and social networks. ID3, a major research nonprofit located in Boston, has also created an implementation called Open Mustard Seed (http://idhypercubed.org/wiki/) focused on how both cloud storage and secure computing can be used.

The New York Times Lab's OpenPaths project is perhaps the most successful of PDS concepts to date. OpenPath users download a client to their smartphone to track their geolocation. This data is then uploaded to a cloud database provided for free where it encrypted and stored. The NYT Lab then provides a set of tools for users to analyze their own location data and facilitates researchers proposing research projects. Most importantly, OpenPaths data is owned solely by the user, can be exported or deleted from the site at any time, and can only be shared with a third parties by users themselves through an active approval mechanism.

As PDS concepts and infrastructure grow in maturity and breadth, they may offer a way to directly interact with individuals who wish to donate their data for the public good.

6. Opportunities and Obstacles for Personal Health Data Research

So what has been learned in this project about whether and how the Robert Wood Johnson Foundation and others should seek to advance research on personal data for the public good? Overall, there is considerable enthusiasm about the potential for research in this area and the barriers appear to be surmountable. A new PHD research ecosystem may well be emerging, and there is an opportunity to make the most of this momentum, while paying attention to prevent unintended adverse consequences of this movement. (Figure 10)

Individuals who generate the data are willing to see it used for research as long as the data are handled anonymously and that it is used for legitimate purposes to advance knowledge in the fields related to PHD such as public health, health care, computer science and social and behavioral science. Researchers see value in these kinds of data, and many consider it to be of equal quality and importance to data from existing high-quality clinical or public health data sources. Researchers think these data can answer important research questions, and a large number see no insurmountable barriers to its use. Most representatives from the companies contacted expressed interest in partnering with researchers, but their responses were more varied. Some small venture-backed start-ups are already viewing the data that they collect as part of their core business and think researchers could add value to it. One large company with millions of users expressed interest "as long as the conditions were right" and there was assurance that the data could be handled in ways that wouldn't put them in jeopardy of a breach of trust with their customers.

However, several issues emerged in this study as potential obstacles to progress if the field of research on personal health data for the public good is to grow. Taken another way, these are opportunities for further analysis, stakeholder engagement and understanding, and policy-level efforts. While some of these issues overlap with one another, the following attempts to categorize them as thematic areas that could be addressed.

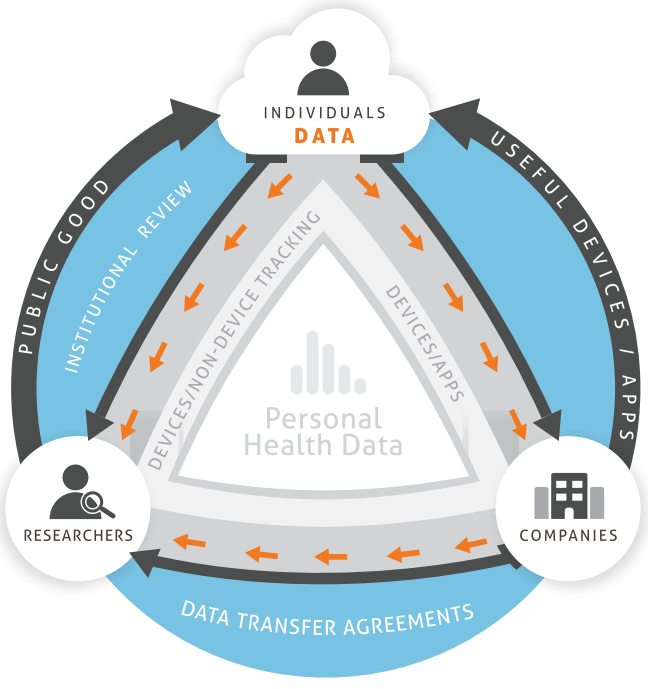


Figure 10. **Personal Health Data Research Ecosystem**

6.1 Data Ownership

Important differences exist with respect to how individuals and companies view ownership of personally generated health data. As noted earlier, terms of use agreements that are signed before use of many personal data technologies typically state that the company providing the technology either fully owns or has full and complete rights to the data, including the right to repackage and sell datasets to others as long as they have been anonymized. In our survey of individuals, while some didn't care who owned the data they generate, a clear majority wanted to own or at least share ownership of the data with the company. Importantly, many thought that they actually did own these data, even though this is unlikely given the most prevalent types of terms of use agreements. While this difference of opinion doesn't appear to be a major barrier at present to growth of use of self-tracking technologies, it may foreshadow a deeper set of public attitudes that could influence future policy making in this area. Thus, there is an opportunity to better understand these attitudes on the part of all stakeholders. This knowledge could contribute to how new policies that might govern the ownership of PHD are developed and could also inform how these policies are put into practice. These policies may become increasingly important as researchers move to combine PHD with more traditional forms of health data for which ownership issues have been clarified (e.g., EMR data).

One alternative approach to increase ownership and access to PDH would be to create a protocol for data exportability similar to the "blue button" initiative at the federal level. A standard protocol developed with input from key stakeholders could facilitate users of all types, including researchers, as they access to self-tracking and other PHD. This could even support sending a copy of this to an open data commons.

6.2 Data Access for Research

Companies, key informants and others interviewed for this project expressed a wide range of opinions about how self-tracked data is (or is not) shared and used for research. Some companies expressed a strong sentiment that they view the data they capture as a corporate asset, a key part of their business model and thus something they would not likely share. On the other end of the spectrum, other companies expressed a highly open approach, including willingness to widely share de-identified data sets. Individuals, while concerned about maintaining their privacy, expressed considerable willingness to have their data shared and used by researches. Their main concerns related to sharing the data for marketing and other commercial purposes.

Even when there is a willingness on the part of a company to make PHD available to researchers, accomplishing this can be an arduous task. A few larger companies have an academic liaison whose job is to respond to requests for partnering and determine which ones to respond to and how. But the sense we got in our interviews is that data access issues based purely upon practical constraints could be a barrier to personal health data research. Creating the right contract language, material transfer agreements or other documentation that satisfies both corporate counsel as well as the research partners is challenging. One company representative stated that: "It took six months to develop contract language for us give some of our data to a leading academic institution at no cost." This presents an opportunity to consider whether templates for these sorts of agreements might be helpful to the field, perhaps one developed and endorsed by a joint industry associations-academic research society partnership or similar approach.

Additional approaches that can address this issue appear to be emerging. One is signaled in what we found with one company, SmallStepsLab, whose business model is to serve as an intermediary between a data rich company, in this case Fitbit, and academic researchers via a "preferred status" API held by the company. Researchers pay SmallStepsLab for this access as well as other enhancements that they might want. Another approach is to advance the use of APIs that open data up for research. As noted above several models of APIs exist but it is as yet unclear if best practices have emerged. Perhaps this field can be advanced through a set of recommended specifications for APIs that can be developed through collaborative efforts of company representatives, researchers and organizations such as IEEE. Another approach might be to foster the adoption of language for data use agreements and terms of service that make it easier for companies to respond if a customer desires to make their data available for research. This could allow a researcher interested in PHD to recruit participants

into a study as long as they were willing to ask their PHD company to release their data for study purposes. Developing consensus about terms of use language that supports such requests could also be accomplished by convening interested stakeholders. Finally, the notion of some form of data repository or data commons surfaced in several of our discussions as well as meetings that several on the HDE team participated in during the project. Mechanisms that allow individuals, companies and/or researchers to place PHD in settings for others to access, perhaps like the personal data locker should be explored as a means to facilitate research in this area.

6.3 Privacy

As noted earlier in this report, policies and practices that relate to privacy of personal health information that emerged in the era of medical records, clinical trials and periodic public health surveys may be insufficient at a time when more and more self-generated data relevant to health are being generated. Users of self-tracking technologies and platforms that collect data that can be analyzed for health research may overlook language in the terms of use that indicate that their data can be used to tailor unique services for them. While these data are typically anonymized, as noted earlier, there is a very real risk of revealing a person's identity if two or more sources of person-generated data are combined.

There is an opportunity to engage in the larger set of privacy discussions stimulated by current events including revelations about the NSA's data collection efforts and emerging concerns about corporate tracking more broadly. Policy documents that specifically address recommendations about how to handle privacy issues for PHD might help protect the availability of these forms of data for research aimed at improving the public good. Based on the research conducted for the HDE project we believe that there are a number of critical issues that need to be addressed by a multi-stakeholder community that involve individuals who are self-tracking, companies creating devices and storing data and academic researchers. First, additional research is needed to help unpack and understand user expectations regarding the privacy of their self-tracking data. This understanding can then help inform conversations regarding establishing norms of use. Second, there is a need to develop appropriate education and outreach materials to help in discussions about the realities and challenges of digital anonymity. Third, tools need to be developed to enhance user control of data, awareness of sharing, and notification of findings derived from the use. These controls are an essential condition for establishing the trust needed to assure that data donation is not one time occurrence.

6.4 Informed Consent & Ethics

Just as these new forms of data raise new questions about data privacy, they create new ones for the ethics of research in this domain, in particular the ethical model we use for assessing the rights of the individuals who donate data and our responsibilities back to them. Most of the current framing of these issues has occurred in a pre-digital era and it is clear that digital data raises unique challenges and opportunities. Much self-tracking data is similar in nature to other types of Internet-based data ranging from blogs to social networks. It would be useful if academics interested in self-tracking and Internet research ethics could come together to discuss existing, newly developed, and future needs for digital human subject data. In a similar fashion, academic self-tracking researchers would benefit from considering new models of consent created to balance the ethical respect for the individual with the scientific need to share data found in large genomic, clinical, and microbiome data sets.

6.5 Research Methods and Data Quality

Several researchers and key informants identified obstacles to progress in PHD research that relate to research methods or to practical issues of conducing this type of research. One of the most common concerns is about data quality, in particular their validity and reliability given the wide variety of sensors and devices that are now in use to capture PHD. Unlike medical devices that undergo a rigorous FDA approval process, consumer-grade self-tracking devices and apps only need pass the test of the marketplace to become widely used. For some types of research such as population-level monitoring of general trends in physical activity, consumer grade pedometers or wearable activity trackers may be acceptable. But if PHD is to be coupled with quality-controlled data (e.g. electronic health record data) and then used to improve health interventions, more will need to

be known about how well PHD devices and apps represent the underlying constructs they measure. A related concern is the potential bias in PHD that derives from who uses personal health devices and who doesn't. Are those from whom these data are collected representative of populations that researchers will be interested in? This presents an opportunity for continued assessment of the characteristics of participants in the PHD ecosystem.

6.6 An Evolving Ecosystem

Finally, we want to emphasize that PHD represents an area in flux. We see this as an opportunity more than an obstacle because the researchers, individuals and companies in this space are in a position to impact the shape of the landscape as it evolves. One area of significant change will be in the area of self-tracking technologies themselves. Right now there are a large number of devices on the market and many more in development. We expect that some of the issues researchers highlighted around the validity of the data and lack of standardization will be addressed as the consumer health device, apps and services market matures. We also expect that as policies are developed, laws are written, and standard practices emerge, some of the uncertainty around ownership, privacy, and ethics will lessen.

Creative solutions must be found that allow individual rights to be respected while providing access to high-quality and relevant data for research, that balance open science with intellectual property, and that enable productive and mutually beneficial collaborations between the private sector and the academy. There is a great deal of experimentation taking place working toward these goals. We are optimistic that the public good can be served by these advances, but we also believe that there is work to be done to ensure that policy, legal, and technological developments enhance the potential to generate knowledge out of personal health data, and ultimately, improve public health and wellbeing.

7. Annotated Bibliography

The accompanying annotated bibliography provides material for introducing key concepts to the lay reader as well as providing in-depth discussion and examples of research. It includes many of the citations in this report with other additions and is intended to evolve other time as new resources are identified. It includes scientific journal articles and white papers, as well as articles from popular media.

Articles from scientific journals are divided into three groups. Articles in the first group present findings from individual studies based on personal data, though not necessarily health data. These articles were selected for interesting features such as obtaining large datasets from companies or recording device data during naturalistic behavior. They serve as examples of the insights that can be gained using these large, personal datasets. The second group of articles has similar features to the above, but focuses on studies based on websites/platforms that were created with the expressed purpose of fulfilling health research. To date, this includes 23andMe, PatientsLikeMe, Personal Genome Project and MedHelp. The third group of references does not contain individual studies, but consists of reviews, editorials and white papers that discuss high-level concepts such as privacy, data access, consent and self-tracking.

The fourth group of citations consists of sources from popular media. These sources are valuable because this field directly depends on individuals who are outside of academia, and they have the potential to characterize or even guide public opinion. Public opinion is particularly relevant to this field as public involvement and trust are fundamental to building these data exchanges.

The area of genomic research is both highly relevant and expansive in content. While many of the references in the bibliography relate to genomic research, a comprehensive review is beyond the scope of this project.

Appendices

See these appendices at http://hdexplore.calit.net/report:

- Copies of survey instruments
- Full Annotated Bibliography

References

Ahmed, A.-K. (2013). With its HeLa genome agreement, the NIH embraces a expansive definition of familial consent in genetics. Retrieved from http://www.michaeleisen.org/blog/?p=1417

Arnason, V. (2004). Coding and consent: moral challenges of the database project in Iceland. *Bioethics*, *18*, 27–49. doi:10.1111/j.1467-8519.2004.00377.x

Asprey, D. (2013). Is Your Self-Monitoring Data Protected by Law. *The Bulletproof Executive BLog*. Retrieved from http://www.bulletproofexec.com/is-your-self-monitoring-data-protected-by-the-constitution/

Ayers, J. W., Althouse, B. M., Allem, J.-P., Childers, M. A., Zafar, W., Latkin, C., … Brownstein, J. S. (2012). Novel surveillance of psychological distress during the great recession. *Journal of Affective Disorders*, *142*(1-3), 323–330. doi:10.1016/j.jad.2012.05.005

Ayers, J. W., Althouse, B. M., Allem, J.-P., Rosenquist, J. N., & Ford, D. E. (2013). Seasonality in Seeking Mental Health Information on Google. *American Journal of Preventive Medicine*, *44*, 520–525. doi:http://dx.doi.org/10.1016/j.amepre.2013.01.012

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*, 17–30. doi:10.1007/s00799-007-0022-9

Brothers, K. B., & Clayton, E. W. (2010). "Human non-subjects research": privacy and compliance. *The American journal of bioethics : AJOB*, *10*, 15–17. doi:10.1080/15265161.2010.492891

Buchanan, E. A. (2010). Internet Research Ethics and IRBs. Chicago: OHRP Research Forum. Retrieved from http://www.hhs.gov/ohrp/sachrp/mtgings/mtg07-10/buchanan20100721.ppt

Buchanan, E. A., & Zimmer, M. (2012). Internet Research Ethics Note: Compared to the AoIR Ethics Guide (http://aoir.org/documents/ethics-guide/), this article intends to be descriptive in nature, not for providing specific guidance. *Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/entries/ethics-internet-research

Clarke, M., Bogia, D., Hassing, K., Steubesand, L., Chan, T., & Ayyagari, D. (2007). Developing a Standard for Personal Health Devices based on 11073. In *29th Annual International Conference of the IEEE* (pp. 6175–7). Lyon. doi:10.1109/IEMBS.2007.4353764

Conger, K. (2012). BIG DATA: What it means for our health and the future of medical research. *Special Report*. Retrieved from http://stanmed.stanford.edu/2012summer/article1.html

Consent to Research. (2013). Retrieved from http://weconsent.us

Cook, D. J., Thompson, J. E., Prinsen, S. K., Dearani, J. a., & Deschamps, C. (2013). Functional Recovery in the Elderly After Major Surgery: Assessment of Mobility Recovery Using Wireless Technology. *The Annals of Thoracic Surgery*, *96*, 1057–1061. doi:10.1016/j.athoracsur.2013.05.092

Davies, S. (2013). Sensor Innovations Driving the Digital Health Revolution. *Bionic.ly*.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376. doi:10.1038/srep01376

Do, C. B., Tung, J. Y., Dorfman, E., Kiefer, A. K., Drabant, E. M., Francke, U., … Eriksson, N. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS genetics*, *7*, e1002141. doi:10.1371/journal.pgen.1002141

Drew, B. T., Gazis, R., Cabezas, P., Swithers, K. S., Deng, J., Rodriguez, R., … Soltis, D. E. (2013). Lost branches on the tree of life. *PLoS Biology*, *11*, e1001636. doi:10.1371/journal.pbio.1001636

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor. Retrieved from http://hdl.handle.net/2027.42/97552

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., … Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS genetics*, *6*, e1000993. doi:10.1371/journal.pgen.1000993

Fox, S., & Duggan, M. (2013). Tracking for Health. *Pew Internet*. doi:10.1001/jamainternmed.2013.1221.2.

Gibson, G., & Copenhaver, G. P. (2010). Consent and Internet-Enabled Human Genomics. *PLoS genetics2*, (June 24). doi:10.1371/journal.pgen.1000965

Glass, T. A., & McAtee, M. J. (2006). Behavioral science at the crossroads in public health: extending horizons, envisioning the future. *Social science & medicine*, *62*, 1650–1671. doi:10.1016/j.socscimed.2005.08.044

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*, 779–782. doi:10.1038/nature06958

Gross, R., & Acquisti, A. (2005). Information Revelation and Privacy in Online Social Networks. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society* (pp. 71–80). doi:10.1145/1102199.1102214

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science (New York, N.Y.)*, *339*, 321–4. doi:10.1126/science.1229566

Hagan, J., & Kutryb, M. (2009). Internet forums track patients' IOL concerns. *Rev Opthalmol*, *16*, 52–5.

Hill, A. B. (1965). The environment and disease: association or causation? *Proc R Soc Med*, *58*, 295–300.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., … Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, *4*, e1000167. doi:10.1371/journal.pgen.1000167

Huberman, B. A. (2012). Sociology of science: Big data deserve a bigger audience. *Nature*. doi:10.1038/482308d

Krumme, C., Llorente, A., Cebrian, M., Pentland, A. S., & Moro, E. (2013). The predictability of consumer visitation patterns. *Scientific reports*, *3*, 1645. doi:10.1038/srep01645

Lane, N. D., Miluzzo, E., Lu, H. L. H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine*, *48*. doi:10.1109/MCOM.2010.5560598

Lane, N. D., Xu, Y., Lu, H., Campbell, A. T., Choudhury, T., & Eisenman, S. B. (2011). Exploiting Social Networks for Large-Scale Human Behavior Modeling. *IEEE Pervasive Computing*, *10*, 45–53. doi:10.1109/MPRV.2011.70

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science*, *323*, 721–723. doi:10.1126/science.1167742.Life

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. In *Proceedings of the 28th international conference on Human factors in computing systems CHI 10* (p. 557). doi:10.1145/1753326.1753409

Li, I., Dey, A., & Forlizzi, J. (2011). Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 405–414). doi:10.1145/2030112.2030166

Lunshof, J. E., Chadwick, R., Vorhaus, D. B., & Church, G. M. (2008). From genetic privacy to open consent. *Nature reviews. Genetics*, *9*, 406–411. doi:10.1038/nrg2360

Madan, A., Cebrian, M., Lazer, D., & Pentland, A. (2010). Social sensing for epidemiological behavior change. *Access*, 291–300. doi:10.1145/1864349.1864394

Madan, A., Cebrian, M., Moturu, S., Farrahi, K., & Pentland, A. "Sandy." (2012). Sensing the "Health State" of a Community. *IEEE Pervasive Computing*, *11*, 36–45. doi:10.1109/MPRV.2011.79

Madan, A., Moturu, S., Lazer, D., & Pentland, A. (2010). Social Sensing : Obesity , Unhealthy Eating and Exercise in Face-to-Face Networks. In *Wireless Health 2010* (pp. 104–110). doi:10.1145/1921081.1921094

Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013). *Teens, Social Media, and Privacy* (pp. 1–107). Washington DC. Retrieved from http://pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., … Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, *339*, 823–6. doi:10.1126/science.1232033

Markoff, J. (2012). Big Data Troves Stay Forbidden to Social Scientists. *New York Times*. Retrieved from http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html?_r=1&&pagewanted=print

Mont, M. C., Sharma, V., Pearson, S., Saeed, R., & Filz, M. (2011). *Technical Architecture Arising from the Third Case Study*. Retrieved from http://www.encore-project.info/deliverables_material/D2_3_EnCoRe_Architecture_V1.0.pdf

Moturu, S. T., Khayal, I., Aharony, N., Pan, W., & Pentland, A. (Sandy). (2011). Sleep, mood and sociability in a healthy population. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5267–5270. doi:10.1109/IEMBS.2011.6091303

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (NCPHS). (1979). *Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Federal Register* (Vol. 44, pp. 23192–23197).

National Human Genome Research Institute (NHGRI), N. (2010). *The Genetic Information Nondiscrimination Act (GINA)*. Retrieved from http://report.nih.gov/nihfactsheets/ViewFactSheet.aspx?csid=81

Neff, G. (2013). Why big data won't cure us. *Big Data*, *1*(3), 117–123. doi:10.1089/big.2013.0029

Niederdeppe, J., & Frosch, D. L. (2009). News coverage and sales of products with trans fat: effects before and after changes in federal labeling policy. *American journal of preventive medicine*, *36*, 395–401. doi:10.1016/j.amepre.2009.01.023

Nissenbaum, H. (2011). A contextual approach to online privacy. *Daedalus, the Journal of the American Academy of Sciences and Arts*, *Fall*, 32–48.

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, *41*, 100–126. doi:10.1111/j.1745-6606.2006.00070.x

Palfrey, J., & Gasser, U. (2008). *Born Digital: Understanding the First Generation of Digital Natives. Hedgehog Review* (Vol. 198, p. 288). doi:10.1097/NMD.0b013e3181cc549e

Pellegrini, C. A., Verba, S. D., Otto, A. D., Helsel, D. L., Davis, K. K., & Jakicic, J. M. (2012). The Comparison of a Technology-Based System and an In-Person Behavioral Weight Loss Intervention. *Obesity*, *20*, 356–363. doi:10.1038/oby.2011.13

Pentland, A., Lazer, D., Brewer, D., & Heibeck, T. (2009). Using reality mining to improve public health and medicine. *Studies in health technology and informatics*, *149*, 93–102. doi:10.3233/978-1-60750-050-6-93

Rabinow, P. (1999). Artificiality and enlightenment: from sociobiology to biosociality. In M. Biagioli (Ed.), *The Science Studies Reader* (pp. 407–17). London: Routledge.

Rainie, L., Kiesler, S., Kang, R., & Madden, M. (2013). Anonymity, Privacy, and Security Online. *Pew Research Center*. Pew Research. Retrieved from http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx

Rouse, M. (2010). HIPAA (Health Insurance Portability and Accountability Act). *SearchDataManagement*. Retrieved from http://searchdatamanagement.techtarget.com/definition/HIPAA

Shoemaker, R., Deng, J., Wang, W., & Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome research*, *20*, 883–889. doi:10.1101/gr.104695.109

Singer, N. (2013, March 30). An American Quilt of Privacy Laws, Incomplete. *New York Times*.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS Quarterly*, *35*, 989–1016. doi:10.1126/science.1103618

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science (New York, N.Y.)*, *327*, 1018–1021. doi:10.1126/science.1177170

Sung, A., Marc, C., & Pentland, A. (2005). *Objective physiological and behavioral measure for tracking depression*.

Sweeney, L., Abu, A., & Winn, J. (2013). Identifying Participants in the Personal Genome Project by Name. *SSRN Electronic Journal*, 1–4. doi:10.2139/ssrn.2257732

Turow, J. (2011). *The Daily You: How the new advertising industry is defining your identity and your worth* (p. 256). Yale University Press.

Windley, P. (2010). Essential Characteristics of a Personal Data Store. *Technometria Blog*. Retrieved from http://www.windley.com/archives/2010/11/essential_characteristics_of_a_personal_data_store.shtml

World Economic Forum. (2011). *Personal Data : The Emergence of a New Asset Class An Initiative of the World Economic Forum. Forum American Bar Association* (pp. 1–40). Retrieved from http://www.weforum.org/reports/personal-data-emergence-new-asset-class

Zimmer, C. (2013, August 7). A Family Consents to a Medical Gift, 62 Years Later. *New York Times*

Acknowledgements

We thank the many individuals, researchers, company representatives and key informants who shared with us their perspectives on personal health data.

Health Data Exploration Project National Advisory Board

Linda Avey, Co-founder, 23andMe and Curious, Inc.
Hugo Campos, Patient Advocate, San Francisco
Robert M. Kaplan, PhD, National Institutes of Health
Sendhil Mullainathan, PhD, Harvard University
Tim O'Reilly, O'Reilly Media
Larry Smarr, PhD, Director, Calit2
Martha Wofford, Aetna
Gary Wolf, Co-Founder, Quantified Self Labs

Robert Wood Johnson Foundation

Stephen Downs, Chief Technology and Information Officer
Lori Melichar, PhD, MA, Senior Program Officer

Health Data Exploration Project

Project Director: Kevin Patrick, MD, MS
Professor, Family and Preventive Medicine, UCSD
Director, Center for Wireless and Population Health Systems, Calit2

Project Co-Director, Jerry Sheehan, MA
Chief of Staff, Calit2

Investigators

Matthew Bietz, PhD, Project Scientist, UC Irvine
Judith Gregory, PhD, Adjunct Professor, UC Irvine
Scout Calvert, PhD, Project Scientist, UC Irvine
Ramesh Rao, PhD, Director, Calit2/UCSD

Researchers

Mike Claffey, PhD Student, UCSD
Alexandra Hubenko, MBA, Program Manager

Communications

Tiffany Fox, Calit2
Jemma Weymouth, Burness Communications

SUGGESTED CITATION:
Personal Data for the Public Good: New Opportunities to Enrich Understanding of Individual and Population Health. 2014. Health Data Exploration Project. Calit2, UC Irvine and UC San Diego.

Supported by a grant from the Robert Wood Johnson Foundation


Health Data Exploration project