CITIZEN SCIENCE:
THEORY AND PRACTICE

Bietz, M, et al. 2019. Data Donation as a Model for Citizen Science Health Research. *Citizen Science: Theory and Practice*, 4(1): 6, pp. 1–11. DOI: https://doi.org/10.5334/cstp.178

ESSAY

Data Donation as a Model for Citizen Science Health Research

Matthew Bietz*, Kevin Patrick† and Cinnamon Bloss†

New computational and sensing innovations, coupled with increasingly affordable access to consumer health technologies, allow individuals to generate personal health information that they are then able to submit to a shared archive or repository. This paper presents data donation as a model for health-focused citizen science, with special attention to the ethical challenges and opportunities that this model presents. We also highlight some existing data donation projects curated by citizen scientists. After describing data donation in more detail, including its relationship to movements like the Quantified Self and research in personalized medicine, we report findings from the Health Data Exploration (HDE) Project's second annual Network Meeting, which was focused on data donation. These findings include identification of four challenges for the ethical conduct of health-focused data donation research: Participant protection, representativeness, incentives to participate, and governance. We use these insights as a springboard for further discussion of specific issues, pointing both to the current state of the field and our suggestions about potential pathways for addressing some of the challenges.

Keywords: citizen science; data donation; informed consent; data access; privacy; data governance

Introduction

The field of citizen science is showing an increasing interest in the domain of human health and wellness, which have not been as well represented as topics like non-human biology, ecology, earth sciences, and astronomy. The expertise, resources, and technologies for the production of health-related research have been sequestered within the professional, regulatory, and market frameworks of clinical medicine and public health. However, movements like the Quantified Self have encouraged individuals to investigate their own bodies, behaviors, and conditions, often using new computational and sensing technologies. Access to medical-grade and "prosumer" health equipment and services (like wearable devices that measure vital signs, or direct-to-consumer genetic testing) is becoming easier and less expensive. The Internet can be a platform for sharing expertise, collating and managing data, and the collective conduct of science. Opportunities abound for health-focused citizen science.

Data donation is receiving significant attention as a form of citizen participation in health-related research. In a data donation study, participants create a federated dataset by submitting their own personal health data to a shared archive or repository. In some cases, the data may have been collected by the participants themselves. The data can be generated from digital technologies such as wearable devices, traces of online activity such as social media posts, or a patient's medical tests or electronic health records. The key point is that the dataset is created collaboratively by the people who are represented within it.

This paper explores data donation as a model for citizen science research in health, with special attention to the ethical challenges and opportunities that this model presents. We also highlight some existing data donation projects curated by citizen scientists. After describing data donation in some detail, including its relationship to movements like the Quantified Self and research in personalized medicine, we report findings from the Health Data Exploration (HDE) Project's second annual Network Meeting, which was focused on data donation. We use insights from this session as a springboard for further discussion of specific issues, pointing both to the current state of the field and our suggestions about potential pathways for addressing some of the challenges.

A Model for Citizen Science Health Research

The focus of this paper is on the concept of data donation as a form of public participation in health research. At the outset, it is important to define what is meant by data donation in this context. Data donation research is *research in which people voluntarily contribute their own personal data that was generated for a different purpose to a collective dataset.* In the context of citizen science

* University of California, Irvine, US
† University of California, San Diego, US
Corresponding author: Cinnamon Bloss (cbloss@ucsd.edu)

and health research, these data may directly or indirectly contribute to an understanding of humans, and they are contributed by the individuals to whom the data refer.

Some of the terms in this definition require further explanation and clarification. First, we must clarify what we mean by "voluntarily contribute." In this case, we are focusing on those instances in which a "data subject" makes a clear choice to allow data about themselves to be used in a research study. In other words, participants "opt in" to the research. So, for example, when individuals have their DNA analyzed by the personal genomics company 23andMe, the company then presents them with a free choice of whether to contribute their genetic data for use in research studies (23andMe 2018). We consider this to be included in the definition of data donation research. On the other hand, some companies compel data sharing in their terms of service, and use data collected from users for internal research, to drive advertising, or to sell to third parties. This does not meet the criteria of voluntary contribution and is not considered data donation research. For example, a fitness device company selling users' data to a third party for research would not be considered data donation research. On the other hand, if a user of that device downloaded their data and contributed it for research on their own, this *would* be considered data donation.

Another important aspect of our definition is that the data that get donated are often originally generated for purposes other than the research study itself. Imagine, for example, someone who for years has worn a fitness tracker to help understand their wellness and used a GPS-enabled mapping service on their phone to help them navigate when they drive. This GPS tracker also incidentally captures data from the user when they exercise. At some point, they discover a data donation project that is about understanding the impact of location on exercise patterns and decide to contribute their activity level and GPS-location history data from which the researchers can extract locations of episodes of exercise. The research represents a secondary use of these data. In this example the data may be generated intentionally (as with a fitness tracker), but also can be generated passively as a byproduct of other activities (as with location history being a byproduct of GPS navigation tracking). Even when the data collection is closely related to the research itself, there is often an individual benefit from the data collection that is separate from the study purposes. For example, personal genomics companies like 23andMe may provide opportunities to contribute genetic data for research, but the initial impetus for individuals to get genetic sequencing is often to understand their health or ancestry.

The development of a collective dataset can allow individuals to compare themselves to others and can yield population-level generalizations. The form of the dataset and the mode of integration are highly diverse. Some projects may federate datasets in a way that focuses on each individual's particular story (like a collection of n-of-1 studies), while others may create new comprehensive and integrative databases that allow for comparisons across specific variables. Similarly, some projects may develop databases that are open to the public and carry few if any restrictions, while others may place tight restrictions on access and use.

It is also important to note some areas that are not covered by our definition of data donation research. For example, we do not make a distinction about whether the research is for-profit or not-for-profit. We do not make a distinction about the scope or intent of the research itself. We also do not make a distinction about how or by whom the data are collected; that is, we do not distinguish between a participant taking measurements manually and writing them in a personal journal versus using a commercial device that automatically uploads personal sensor data to a company's server, as long as the user of such devices or services has the ability to access and share the data with a research project.

Throughout this paper we use a number of terms to refer to the people involved in research. Sometimes we use the general labels of "people" or "individuals." Other terms reflect specific roles and positions within the spectrum of research. A "subject" is the object of research – the things being studied. "Human subject" is the term most associated with academic research ethics, and simply means that the subject of research is human (as opposed to being rocks or rabbits, for example). The term "data subject" is used to highlight the fact that sometimes research is not done on the human directly (like in medical trials or psychology experiments), but only on the data traces that have been generated from their online activity. The term "participant" is more specific than "subject," referring to an individual who not only is a human subject of research but also has some level of intentional involvement in the research study. It is possible (although often unethical) to study individuals without their participation. At a minimum, "participation" implies that the individual has made an informed and autonomous decision to become a subject of the research, and this is how the term is used in traditional research ethics conversations. In the context of citizen science, it is hoped that participant involvement will go further, perhaps to include activities like shaping research plans, managing and analyzing data, or writing up results. Finally, the term "patients" includes people who are involved in some sort of medical care. Patients are an important category within health research because a large amount of data is often generated from medical procedures that could be of interest for health research. However, not all patients are research subjects, and not all research subjects are patients.

Data donation could serve as a productive model for citizen science research in health and human behavior. However, not all data donation research is necessarily citizen science. For example, data donation is also a model that can be used in traditional academic and corporate research projects. The definition of data donation research outlined above suggests that these projects represent, at a minimum, contributory forms of public participation in scientific research (Bonney et al. 2009; Shirk et al. 2012). Bonney et al. outline three levels of public participation in research: *Contributory* projects where members of the public primarily contribute data; *collaborative* projects

where members of the public may assist with research design, analysis, or dissemination; and *co-created* projects in which members of the public and scientists work together on a more equal footing. Data donation could be used in all of these levels of participation. When brought into a citizen science context, data donation has the potential to empower individuals to reuse and repurpose data that have been collected within a commercial context (e.g., through apps or devices) for projects that they decide are important.

Vignette: A (Fictional) Data Donation Citizen Science Project

This vignette provides a fictional example that describes how an individual ("Jada") might experience data donation in the course of participating in a citizen science project.

In a discussion with one of her friends, Jada hears about a citizen science project investigating whether there is a genetic component that could explain sleep quality. Jada visits the project website and discovers that she meets the minimum qualifications: She has had genetic testing conducted by 23andMe and also uses a sleep tracking device. She clicks on the "Join the Project" link, reads through some information about the project, agrees to the research consent form, and fills out a personal profile. She is then taken to a page where she can share data for the study. First Jada needs to supply her genetic data. 23andMe makes raw data available to participants in downloadable files, so the project website provides instructions on how to retrieve those data and then upload them to the study databank. After finishing the upload, Jada needs to provide her sleep data. Because she uses a popular wearable fitness device that tracks her sleep, and because the company that makes the device also provides an API (application programming interface) that allows for a direct data transfer, donating these data is a simpler process. Jada clicks the "Link My Device" button, which opens a new window asking for her device username and password. After signing in, she then confirms that she wants to allow the project to access her device data and clicks OK. She can share other kinds of data that might be useful using the same processes. She also decides to share her level of physical activity (collected by the same fitness device), a food diary (that she keeps using a smartphone app), and some medical test results (that she downloads from her healthcare provider's electronic health record portal). She also keeps track of her mood on a daily basis in a spreadsheet and decides to upload that as well. Depending on the kind of data, each file would automatically or manually have identifying information removed before being integrated into the dataset collected from all the project members. Jada is looking forward to joining in the online project discussions and helping to analyze these data as the repository grows in size.

Data Donation, Human Subjects, and Health Research

When we are discussing data donation for health research, it is important to first note that these data are about people. Some data are obviously "health data" (e.g., a log of an individual's blood pressure or glucose levels), while other data may not be so obviously human related. One potential source of data that could inform studies of human health is the "Internet of Things" (IoT), including "smart" devices that we might have in our homes. The Nest Thermostat, for example, uses a variety of sensors and artificial intelligence to optimize heating and cooling in the home while conserving energy. To do this, the device collects data not only about temperature but also about humidity, light levels, and movement in the house (Nest Labs 2018). These data are ostensibly about the house but could potentially be used to study such things as sleep patterns or level of individual activity.

Second, as a citizen science project, these data are being contributed by the people they are about. This is especially important when so many data sources today involve consumer devices and commercial organizations. Thus, excluded here are projects that involve obtaining a dataset of many users' data directly from a company. In citizen science health research, individuals who are the subjects are not simply allowing data to be collected about them but are actively participating in the donation and curation of the data.

Interestingly, the roots of data donation research can be found not so much in the realm of other citizen science projects but instead in the practices of traditional biology and genetics research. Ankeny and Leonelli (2015) trace the origins of data donation as a model for scientific research to changes in scientific publication in the genomic era. In particular, while early genomic databases (e.g., GenBank [NCBI n.d.]) began by harvesting genomic data from published scientific literature, they quickly found that they simply could not keep up with the rate at which genomic data were being produced. Instead, they changed procedures so that scientists could submit their own data directly to the databases (Hilgartner 1995). Around the same time, journals and funders began requiring that scientists "donate" their data to public databases (McCain 1995; Contreras 2011). In more recent years, the push toward open data sharing has become more pervasive in academic research broadly, but comprehensive data donation to shared databases is still most pronounced in the genomic sciences.

Thus, it is not particularly surprising that some of the first large-scale citizen data donation projects revolved around genomic data. Given that those involved had worked with collaborative databases populated with data donated by academic researchers, creating databases of publicly donated data perhaps felt familiar to them. One of the first large-scale moves in the direction of citizen data donation was the Personal Genome Project (PGP), in which participants would provide samples for whole-genome sequencing (OHF n.d.b.). Participants' sequence data would be donated to a public database, which could be used by any researchers and for any purpose. The American Gut Project (AGP) took a similar approach, but with gut microbiome sequencing and data sharing (American Gut n.d.). The AGP also used a crowdfunding model in which participants paid a fee to join the project and have their samples sequenced. Unlike many

traditional genomic research projects, both the PGP and the AGP participants are provided with full access to their own data.

These genomic examples also demonstrate that the data donation model described in the previous section may not be fully realized in all projects that use it. For example, data donation projects typically involve data where the research is not the primary purpose for data collection. In such cases, the research may have been a primary reason for participation. The kind of sequencing conducted in the PGP and AGP was available only because the data would be used for research, and contributing to open science was a key selling point for the projects (and the science has been quite successful, see McDonald et al. [2018]). On the other hand, both projects also promised that participants would gain access to their own results that could (potentially) be used to better understand themselves, and individuals would also be free to donate the data generated to other research projects. It is not possible to fully untangle participants' motivations in these projects. However, both projects have been characterized as "data donation" rather than traditional participant recruitment (Harvard PGP n.d.; Liu et al. 2017). It is also important to note that while the PGP and AGP may not in themselves completely conform to the definition of data donation research that we have proposed here, these projects have both enabled further data donation by providing participants with personal genomic data that they can donate to other projects as well.

In addition to these projects focused on genomic data, there is also increased interest in using new forms of personal data in health research (Bietz et al. 2015). Most health research data were traditionally generated from clinical or epidemiological studies. New wearable and pervasive technologies (like IoT or "smart" devices), however, generate data that have potential value for health research. Consumer-level wearable devices collect data like activity levels and vital signs (heart rate, blood pressure, galvanic skin response) and transmit those data wirelessly through apps on users' smart phones. Other devices in our homes and personal spaces can provide streams of data about individuals (e.g., sleep quality), their lifestyles (technology use, eating habits), and other environmental factors (e.g., air quality). The digital traces that we leave online both intentionally (like social media posts) and unintentionally (web browsing histories) may reveal both behaviors and states of mind. The HDE Project, described in more detail below focuses on the feasibility and opportunities for using these new forms of data in health research (**Figure 1**).

This work is also aligned with the goals of the Quantified Self (QS) movement (QS Labs 2015; Wolf 2009). QS brings together people who are interested in using data collection and technology to quantify and analyze various aspects of their bodies and lives. Individuals participate through local "Meetups" where they discuss their data collection and findings, demonstrate tools and methods, and network with like-minded individuals (Nafus and Sherman 2014). While QS is often focused on individual-level data and self-quantification, there is significant interest within the movement in conducting larger collective citizen-driven research with the kinds of data that individuals collect about themselves (Barrett et al. 2013).

At present, one of the most comprehensive approaches to citizen data donation for health and social research is the Open Humans Project. The Open Humans Foundation was created by some of the same individuals who ran the PGP, and is a non-profit organization funded by grants from the Robert Wood Johnson Foundation, Knight Foundation, and the Shuttleworth Foundation. Open Humans functions as a clearinghouse to support citizen science through data donation:

> Open Humans is a platform that allows you to upload, connect, and privately store your personal data – such as genetic, activity, or social media data. Once you've added data, you can choose to donate it: You might choose to share some publicly, and you can join and contribute to diverse research projects. Thus, we turn the traditional research pipeline on its head: you are at the center and in control of when you share your data. We want to empower you to explore your data – for example, enabling you to analyze your genome or your Twitter data (OHF n.d.a).

At the same time, Open Humans also provides researchers and citizen scientists with "a toolbox to easily create new projects that can efficiently ask an engaged audience of participants to join and contribute" (OHF n.d.a). Open Humans extends the data donation models from the PGP and AGP studies to a much wider set of data types and sources. As a community-driven platform, they also welcome the donation of new tools as well as data. This pushes the data donation model beyond just contribution of data into more co-created and participant-led science.

Human Subjects Research Ethics Practices

As citizen science projects increasingly study humans, this raises a number of ethical considerations, especially for the treatment of research subjects. Academic research has predominantly dealt with the ethical issues surrounding human subjects research by developing institutional mechanisms (often mandated by funders) to ensure that research is conducted ethically. For example, in the United States (US), most academic research is governed by a set of ethical guidelines known as "the Common Rule" (USDHHS n.d.). While the Common Rule is subject to interpretation (and indeed, its operationalization can vary dramatically), it codifies the ethical responsibilities and institutional mechanisms to govern human subjects research in the US. Similar ethics review processes are mandated in other parts of the world, for example, in the European Commission Horizons 2020 funding process (European Commission n.d.).

Data donation as a form of citizen science presents two fundamental challenges to our current set of human subjects ethics practices. First, the Common Rule operates through a logic of institutional research, targeting organizations like universities as the bearers of responsibility for operating in an ethical manner. With these

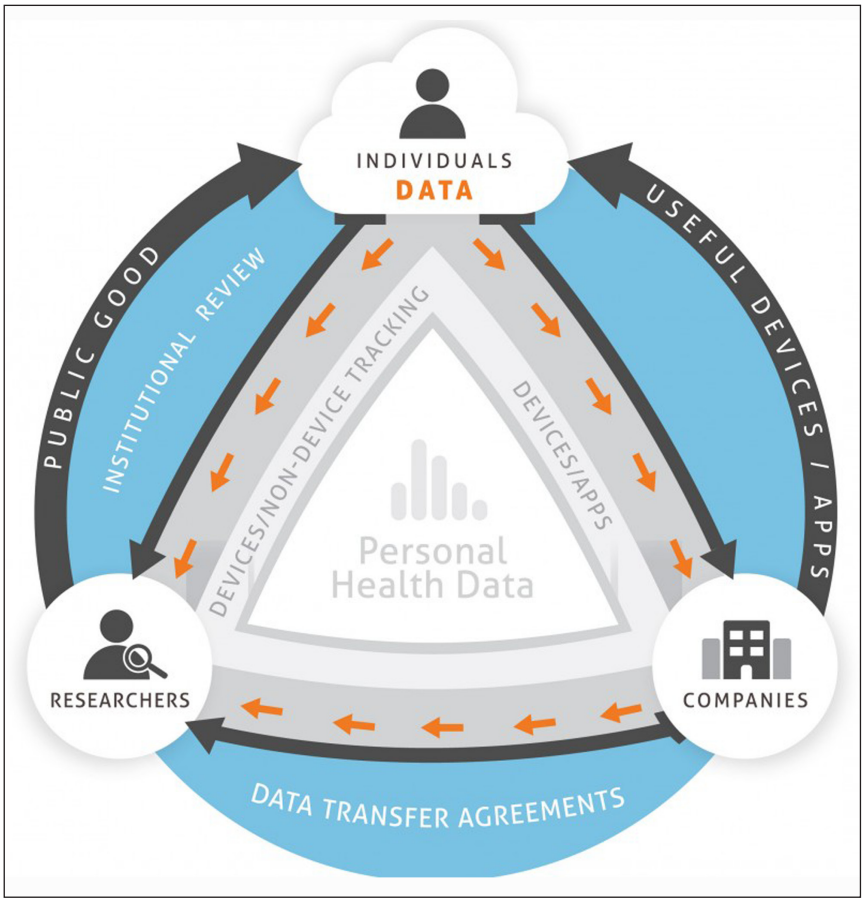

Figure 1: The Health Data Exploration Network brings together innovators in Personal Health Data to catalyze the use of personal data for the public good.

organizations, ethics are managed through a set of institutional practices, most commonly systematic IRB review. Citizen science tends to de-institutionalize or disrupt the traditional institutional research infrastructure, however. Even if projects have some academic affiliation, many of the researchers involved in a citizen science project may not have institutional connections and are thus outside the regulatory regime of the Common Rule. Also, projects that do not have a direct academic affiliation may not be able to draw on the ethical support structures that are in place in academic environments.

Second, traditional human subjects research protections are ill-suited to deal with decentralized, personal-datacentric research. The Common Rule and related guidelines are built on a certain set of conceptions about academic research. For example, they tend to assume that research will be conducted by an identifiable research team, usually distinct from the subjects of the research. That research team is assumed to be directly responsible for conducting any necessary research procedures and for generating the data on which the research is based. Except in extraordinary circumstances, it is assumed that the human subjects of the research should be informed about the research and give their consent to participation before any intervention or data collection is undertaken. In a citizen science data donation context, assumptions like these may not hold true. For example, data donation projects ask participants to submit data about themselves, breaking down the distinction between researcher and subject. In many data donation projects, the machinery of data production is outside the hands of the researchers. It also may have been in operation for a significant time before the research (and associated ethical review) was initiated, and often depends on commercial devices and services that may be resistant to ethical regulation.

Others have begun to address ethical challenges in related areas with a focus on how new forms of pervasive data are being generated at a scale and level of detail that challenge ethical norms and thus generate new concerns for participant protection, data access, and research legitimacy (Rothstein et al. 2015; Vayena et al. 2012; Vayena et al. 2015; Weibel et al. 2017). For example, some have suggested that these new forms of data require a rethinking of traditional ethical review structures within academia (Bloss et al. 2016). Vayena and Tasioulas (2013) suggest a set of three categories of participant-led research for determining appropriate ethical oversight mechanisms. They refer to their first category as "institution-plus," which includes projects that are affiliated with a state-recognized or profit-making institution. Projects in this category have identical oversight obligations as standard research. The second category involves research that has no institutional affiliation but involves more than minimal risk. These projects require external ethics review that may be of a non-standard form; for example, a crowd-sourced review (Swan 2012). The third category is projects

with no affiliations that present no more than minimal risk, thus no formal review is necessary.

While these proposals cover some of the challenges raised by new forms of data and participant-led research, they have not addressed data donation. Citizen science projects that study humans and rely on data donation cannot rely solely on an ethical framework that is both institutionally based and ill-suited for the modalities of this kind of science. To add to this prior work, we next present results from a stakeholder-driven brainstorming session focused on addressing some of these issues in data donation research.

HDE Network Meeting Methods and Results

The HDE Network comprises approximately 300 individuals interested in the opportunities and challenges for using new forms of personal data for health and behavior research. This network was developed by the Health Data Exploration (HDE) Project, a Robert Wood Johnson Foundation-funded research project, which is led by the co-authors (PI Patrick). The network includes academic and corporate researchers, developers of new personal health technologies, citizen scientists, and other interested stakeholders. The HDE Annual Network Meeting convened on May 17, 2016 in San Diego, California with the theme "Enabling Personal Data Donation for Public Good Research." This meeting brought together a diverse group of approximately 150 scientists (academic, corporate, and citizen), developers, and industry partners representing several disciplines including design, public health, bioinformatics, technology, big data, citizen science, and bioethics. As part of the meeting, a brainstorming session was facilitated in which attendees joined one of 10 small breakout discussion groups. Each group had a pre-assigned facilitator and a specific prompt focused on a particular barrier or facilitator to personal health data donation. Participants were free to join whichever group they wished. Participants worked in these breakout groups to evaluate specific themes related to data donation in the ever-evolving context of emerging technologies and big data. Furthermore, participants made recommendations for future actions toward methods for appropriate stewardship and governance of personal health data donation.

The brainstorming discussions were intended to generate a stakeholder-driven set of priorities. In other words, the goal of these discussions was not to generate solutions but instead to identify areas that need to be addressed to enable and improve data donation research. The participants in those discussions are people who have developed expertise in this area.

Four prominent themes emerged from the small group discussions (**Figure 2**):

1. Participant Protections. There is a need to explore new models and procedures for informed consent. The risks associated with donation of personal health data are not well understood. Such data may reveal information about others who may not necessarily be study participants (for example, genetic test results may reveal the presence of a disease in other family members).
2. Representativeness of Data. Datasets compiled from data donation may not be representative of populations. Some groups of people may be more or less likely to donate their data. We need to understand the potential sources of bias in the dataset of a self-selected sample.
3. Incentives for Participation. What are the benefits for participating in data donation research? How do projects build longer-term engagement with participants? Conference attendees were concerned with the implementation of a compelling incentive system that would appeal to a wide range of potential research participants.
4. Governance. Health research has tended to rely heavily on standardization of research practices, methods, and metrics, as well as the scientific, cultural, and ethical norms that guide them. Participants felt that the issue of data heterogeneity and stewardship could not be disentangled from questions about standardization and governance. How should standards be enforced, and what is the role of existing institutions in building, maintaining, and enforcing those standards?

Discussion

This section discusses and extends each of the themes generated by the participants at the HDE Network Meeting and outlines a set of challenges we believe must be addressed if data donation citizen science is to emerge as an important contributor to health research.

Participant Protections

One of the most important aspects of ethical human subjects research is the protection of research participants. There are many well-known examples of unethical research studies that have resulted in physical, psychological, or social harm to participants, including the Tuskegee Syphilis study, the Milgram Experiment, and the Stanford Prison Experiment (CDC 2015; Milgram 1974; Zimbardo 2018). In the US, systematic ethical regulation of research arose in part in response to these ethical lapses. Practices aimed at participant protection tend to focus on two primary areas: Accounting for and minimizing the risks posed by a research study, and obtaining consent from participants who are well informed of those risks. Both aspects are complicated for data donation citizen science.

Specifically, characterizing the risks associated with data donation can be quite difficult. Data that may seem relatively innocuous on their own could be used to make unintended inferences about personal routines or behaviors. Even something as simple as the number of steps walked in a day could reveal whether an employee who took a day off from work was really home sick in bed. Similarly, the risks associated with individual data may increase when compared to other individuals or combined with other sources of data. In one recent case, the release of aggregate data about user locations revealed sensitive information about US military bases (De Mooy 2018).

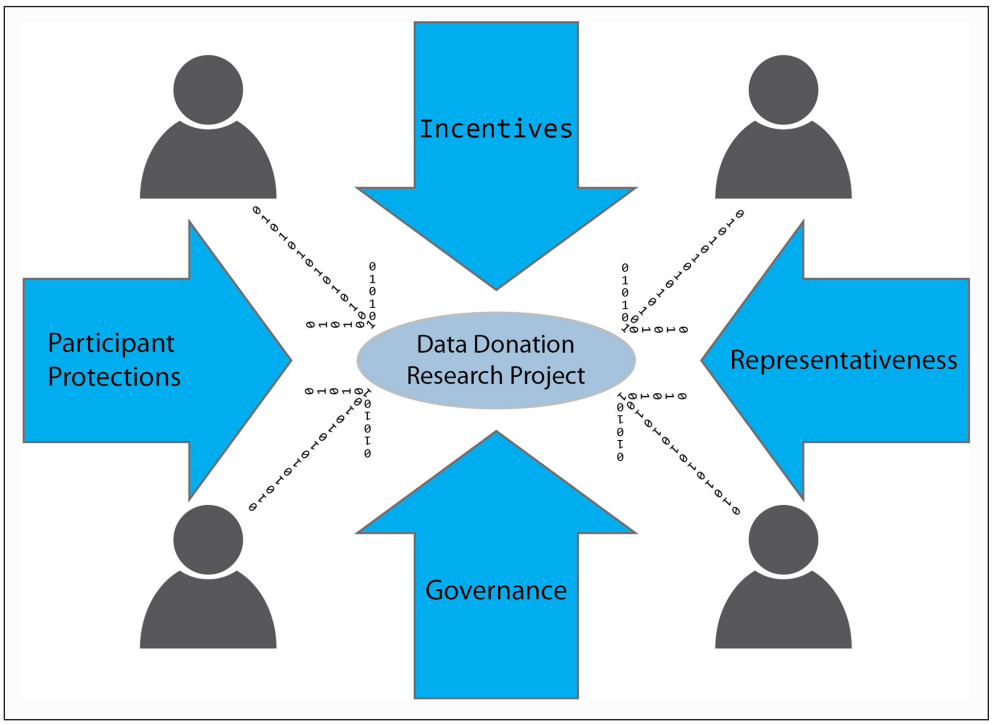

Figure 2: Participants in the Health Data Exploration Network Meeting identified four prominent themes for data donation citizen science.

Data donation projects also may involve the creation of a public or semi-public dataset. For example, the PGP and AGP both create public datasets of genomic data. These data are anonymous, but it may be possible to recover identities from these kinds of data, especially if they are linked to other forms of data (like health records, demographics, or location information). In these cases, it may be impossible at the beginning of the project to say how the data will be used, and thus what risks are present. This concern is evident, for example, in the training and informed consent documentation that the PGP provides to participants. These materials warn of the potential for nefarious uses of the data, including using the PGP database to synthesize DNA evidence to plant at a crime scene. This raises the issue of the role and practice of informed consent for data donation citizen science. There are very practical questions to be addressed about how to manage a consent process that takes place entirely over the Internet and without face-to-face interaction with a study team. There are also deeper questions about what it means to give consent when a dataset is going to be public and can be used for any purpose.

Data donation projects also highlight issues of time and duration with regard to research datasets. While data donation may be a one-time event, some forms of personal data are generated as streams rather than as discrete events. Technologies like APIs allow for direct computational access to data sources and make it possible that the data donation may involve providing access to an ongoing data stream. In these cases, it will be important to consider how long this authorization will remain active and what ability the participant has to revoke access in the future. Researchers have an ethical responsibility to consider and plan for the eventual fate of datasets, either by ensuring that they are destroyed after a certain amount of time or planning for longer-term disposition and curation. Participant-led efforts, especially if they lack institutional partners, may not have the resources to properly manage the data over longer time scales.

Representativeness of Data

One concern with data donation-based research is ensuring that the datasets provide adequate representation of the populations under study. Several factors may result in a biased dataset. Traditional health studies often have recruitment efforts and targets aimed at ensuring that the study participants are a representative sample of a larger population of concern. Recruitment by race, ethnicity, and gender is monitored to prevent the kinds of discrimination in medical research that has occurred in the past (NIH 2018). Data donation faces all of the issues that traditional research faces around issues of recruitment and representation, but also presents several specific challenges to representativeness of the sample.

One challenge is that not everyone has data available to donate. The technologies and practices of data generation are unevenly spread through society. "Billions of people worldwide remain on big data's periphery. Their information is not regularly collected or analyzed, because they do not routinely engage in activities that big data are designed to capture" (Lerman 2013). Data donation often asks individuals to donate data that they have collected themselves, often with the assistance of commercial technologies. But this requires individuals to purchase "smart" technologies like activity monitors, sleep monitors, and bathroom scales, which transmit data over the Internet to databases run by the manufacturer. Users are then able to access their data through a website or mobile app,

sometimes after paying a subscription fee. In other words, even if someone has produced data, financial or other demographic barriers may prevent accessing the information in ways that are necessary to participate in data donation research (Anderson 2017). For example, iPhone users tend to be more affluent, more educated, and more racially and ethnically homogenous than Android mobile phone users (Smith 2013). Donating data may also demand a certain level of technical expertise to access, manipulate, and upload data in appropriate formats. Other concerns, like privacy, may have demographic or cultural features that result in unequal participation in data donation projects, resulting in nonrepresentative samples.

Incentives and Value for Participation

To develop robust data donation projects, it is critical to consider how the participants will relate to the project. Especially if there is a goal of fostering ongoing donation over a longer term, participants need to see value in participating. This is important for recruiting participants, but there is also an ethical dimension: While altruism may be a motivator for some, providing more immediate or tangible benefits may also matter. In some cases, simply providing tools with which to explore or analyze data may be enough. Engaging in more than just donation, for example, in the development of research questions and methods, the analysis of data, or the presentation of results can also be a strong motivator. It is also important to think carefully about how credit (like being listed as an author or recognized as a contributor) should be assigned for data donation studies. Providing appropriate incentive systems can also help to reduce the potential biases in the datasets by appealing to a wider range of potential research participants.

One common way to provide value to participants in data donation research is to provide some form of data analysis, visualization, or other return of research results to participants. Tools developed for the research may be able to provide participants with new insights or understandings of their data. However, returning research results to participants presents several ethical questions. In many cases, the results may be inconsequential. However, it is possible that returning a significant finding (for example, the presence of a harmful genetic mutation), especially without institutional support structures in place, could lead to participant harm (Fabsitz et al. 2010). At the same time, others have argued that it would be an ethical lapse to withhold results from participants (Knoppers et al. 2006). Deciding how to support both the research and participants' own exploration of their data is a key ethical challenge for data donation projects.

The Need for Governance

Participants at the HDE Network Meeting recognized a need to identify and develop appropriate governance mechanisms. One aspect of this governance is, of course, thinking through how to manage the ethical issues that have been discussed above. Vayena and Tasioulas (2013) identified a taxonomy of participant-led research types that could help to identify when existing institutional governance is enough, or when new or adapted procedures need to be developed. The HDE Network Meeting brainstorming session also focused on questions of data heterogeneity and stewardship as an opportunity for community-level governance and standards-making. There is a need for standards and practices to better support flows and controls of personal data. Similarly, stronger data standards would make it easier to integrate data from different sources, but these standards require buy-in from a wide array of stakeholders, including researchers, individuals with personal data (who may or may not be involved in conducting the research), and the companies and others who develop data generation and collection technologies.

Participants in the HDE Network Meeting brought up one model that is common in large-scale research, in which a centralized body or coalition of stakeholders develop and enforce data standards that allow for federation of activities and databases. This approach can be seen, for example, in multi-site cancer studies where data from many different projects and countries are brought together to make stronger claims (Rolland et al. 2017). This model could be appropriate for large-scale data donation studies. Even without a goal of creating larger datasets, however, standardization of data formats can make the development of and participation in data donation studies much easier. Standardization of ethical norms and practices can also be a powerful way to ensure that the project is conducted in an ethical manner. However, it is not clear whether this level of standardization and governance is the best approach for smaller, community-led data donation projects where data federation is less of a concern.

Regardless of the size of the project, governance remains a concern. How will projects decide what data to accept or not accept; what are appropriate methods; what are appropriate questions to ask and uses for the data; and what are appropriate ethical standards for the project and the communities that are involved? Because the citizen scientists are also the research subjects, there is both a greater need and stronger direct incentives for the community to be involved in setting policies and discussing governance matters.

For data donation projects, it will be important to consider issues surrounding the long-term fate of the data collected. This has already become an issue for Data Donors, a former data donation project run by the Wikilife Foundation. The project simply deleted all of its data when it closed (Wikilife Foundation n.d.). Many data donation projects hope to create longer-lived archives, but long-term data retention raises a number of questions: Will individual participants be able to restrict use of their data in the future? How long will data be kept, and how widely will they be distributed? Are any restrictions placed on how the data can be used or what kinds of studies are acceptable? What responsibilities and conditions should be imposed on those who use the data in the future (e.g., with respect to return of research results or notification to the original participants)?

Conclusion

This paper presents data donation as a model for health-focused citizen science and identifies challenges for the ethical conduct of this research: Participant protection, representativeness, incentives to participate, and governance. Returning to the fictional vignette offered above allows consideration of how each of these challenges might apply to the study that Jada joined. For example, it is important to ask how the study protects the sensitive information like genetic data and medical test results that it collects from Jada. In our vignette, Jada seems excited about participating in the study, but it is not clear that initial enthusiasm will be a sufficient motivator to draw enough participants or sustain their engagement over time. Also, Jada seemed to come across the study serendipitously, and she just happened to have the right data available to participate; however, it is important to ask if everyone would have equal access to the study prerequisites, and if not, how unequal access might produce a biased dataset. Finally, it is not clear in the vignette how the project is governed, or what might happen to the project and its data over the long term.

Donating personal data for research remains a relatively new form of health-related citizen science, and as it develops, it will be important to address these ethical concerns. Encouraging developments happening in this space such as the Open Humans platform, which builds consent mechanisms into the data donation process and prompts project creators to create consent materials as part of project setup, will be important. Moreover, new models for consent and governance for citizen science projects are also being explored, such as the "Blood Testers" project, conducted by members of the Quantified Self community focusing on high-frequency self-testing of blood lipids. This group developed a form of "self-consent" to address ethical issues in participant-led research (Quantified Self 2018).

Health data generation is becoming pervasive. Individuals are creating data traces as they use social media and other online services, the IoT, wearable devices, smartphones, and various other environmental and personal sensors. If these data can be made available for research, they could transform the study of human health and behavior. Individual donation of personal data to a collaborative database can enable a new and, with ongoing care, ethical form of citizen science. Moreover, this model of research also has economic implications, and may ultimately offer more cost-effective and sustainable modes for conducting research. Data donation as a form of health-related citizen science is a field in its infancy, however, and all the points that we raise here warrant further study.

Acknowledgements

The authors thank Cynthia Cheung, M.P.H., M.A. and Rasheed Al Kotob for their assistance with aspects of this project. We also thank the attendees of our Health Data Exploration Network Meeting who participated in the workshop that we describe in this paper.

Funding Information

This work was supported by a grant from the Robert Wood Johnson Foundation entitled "The Health Data Exploration (HDE) project" (PI: Patrick, #71693, 2013–2017); a grant from the National Human Genome Research Institute entitled "Impact of Privacy Environments for Personal Health Data on Patients" (PI: Bloss, R01 HG008753, 2015–2018); and a grant from the National Science Foundation entitled "Pervasive Data Ethics for Computational Research" (PI: Bietz, #1704598, 2017–2021).

Competing Interests

The authors have no competing interests to declare.

Author Contributions

MJB made substantial contributions to the acquisition of data for the work; contributed to drafting the work; provided final approval of the version to be published; and agrees to be held accountable for all aspects of the work.

KP made substantial contributions to the acquisition of data for the work; critically revised the work for important intellectual content; provided final approval of the version to be published; and agrees to be held accountable for all aspects of the work.

CB made substantial contributions to the acquisition of data for the work; contributed to drafting the work; provided final approval of the version to be published; and agrees to be held accountable for all aspects of the work.

References

23andMe. 2018. Research – 23andMe. Available at: https://www.23andme.com/research/ [Last accessed 15 October 2018].

American Gut. n.d. American Gut Overview. Available at: http://americangut.org/american-gut-overview/ [Last accessed 15 October 2018].

Anderson, M. 2017. Digital divide persists even as lower-income Americans make gains in tech adoption. Available at: http://www.pewresearch.org/fact-tank/2017/03/22/digital-divide-persists-even-as-lower-income-americans-make-gains-in-tech-adoption/ [Last accessed 15 October 2018].

Ankeny, R and **Leonelli, S.** 2015. Valuing data in postgenomic biology: How data donation and curation practices challenge the scientific publication system. In: Stevens, H and Richardson, S (eds.), *Postgenomics: Perspectives on biology after the genome*, 126–149. Durham, NC: Duke University Press.

Barrett, MA, Humblet, O, Hiatt, RA and **Adler, NE.** 2013. Big data and disease prevention: From quantified self to quantified communities. *Big Data*, 1(3): 168–175. DOI: https://doi.org/10.1089/big.2013.0027

Bietz, MJ, Bloss, CS, Calvert, S, Godino, JG, Gregory, J, Claffey, MP, Sheehan, J and **Patrick, K.** 2015. Opportunities and challenges in the use of personal health data for health research. *Journal of the American Medical Informatics Association*, 23(e1): e42–e48. DOI: https://doi.org/10.1093/jamia/ocv118

Bloss, C, Nebeker, C, Bietz, M, Bae, D, Bigby, B, Devereaux, M, Fowler, J, Waldo, A, Weibel, N, Patrick, K, Klemmer, S and **Melichar, L.** 2016. Reimagining human research protections for 21st century science. *Journal of Medical Internet Research*, 18(12): e329. DOI: https://doi.org/10.2196/jmir.6634

Bonney, R, Ballard, H, Jordan, R, McCallie, E, Phillips, T, Shirk, J and **Wilderman, CC.** 2009. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. *A CAISE Inquiry Group Report*. Available at: https://eric.ed.gov/?id=ED519688 [Last accessed 15 October 2018].

Centers for Disease Control and Prevention. 2015. The Tuskegee Timeline. Available at: https://www.cdc.gov/tuskegee/timeline.htm [Last accessed 15 October 2018].

Contreras, JL. 2011. Bermuda's legacy: Policy, patents, and the design of the genome commons. *Minnesota Journal of Law, Science & Technology*, 12(1): 61–125.

De Mooy, M. 2018. The ethics of design: Unintended (but foreseeable) consequences. Available at: https://cdt.org/blog/the-ethics-of-design-unintended-but-foreseeable-consequences/ [Last accessed 15 October 2018].

European Commission. n.d. Participant Portal H2020 Online Manual: Ethics. Available at: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm [Last accessed 15 October 2018].

Fabsitz, RR, McGuire, A, Sharp, RR, Puggal, M, Beskow, LM, Biesecker, LG, Bookman, E, Burke, W, Burchard, EG, Church, G, Clayton, EW, Eckfeldt, JH, Fernandez, CV, Fisher, R, Fullerton, SM, Gabriel, S, Gachupin, F, James, C, Jarvik, GP, Kittles, R, Leib, JR, O'Donnell, C, O'Rourke, PP, Rodriguez, LL, Schully, SD, Shuldiner, AR, Sze, RKF, Thakuria, JV, Wolf, SM and **Burke, GL.** 2010. Ethical and practical guidelines for reporting genetic research results to study participants: Updated guidelines from an NHLBI working group. *Circulation: Cardiovascular Genetics*, 3(6): 574–580. DOI: https://doi.org/10.1161/CIRCGENETICS.110.958827

Harvard PGP. n.d. The Harvard Personal Genome Project (PGP) – enabling Participant-driven Science [Online]. Cambridge, MA: Harvard Medical School. Available: https://pgp.med.harvard.edu/ [Accessed 14 October 2018].

Hilgartner, S. 1995. Biomolecular databases: New communication regimes for biology? *Science Communication*, 17(2): 240–263. DOI: https://doi.org/10.1177/1075547095017002009

Knoppers, BM, Joly, Y, Simard, J and **Durocher, F.** 2006. The emergence of an ethical duty to disclose genetic research results: International perspectives. *European Journal of Human Genetics*, 14(11): 1170. DOI: https://doi.org/10.1038/sj.ejhg.5201690

Lerman, J. 2013. Big data and its exclusions. *Stanford Law Review Online*, 66: 55. DOI: https://doi.org/10.2139/ssrn.2293765

Liu, Y, Ferreira, D, Goncalves, J, Hosio, S, Pandab, P and **Kostakos, V.** 2017. Donating context data to science: The effects of social signals and perceptions on action-taking. *Interacting with Computers*, 29(2): 132–146. DOI: https://doi.org/10.1093/iwc/iww013

McCain, KW. 1995. Mandating sharing: Journal policies in the natural sciences. *Science Communication*, 16(4): 403–431. DOI: https://doi.org/10.1177/1075547095016004003

McDonald, D, Hyde, E, Debelius, JW, Morton, JT, Gonzalez, A, Ackermann, G, Aksenov, AA, Behsaz, B, Brennan, C, Chen, Y, DeRight Goldasich, L, Dorrestein, PC, Dunn, RR, Fahimipour, AK, Gaffney, J, Gilbert, JA, Gogul, G, Green, JL, Hugenholtz, P, Humphrey, G, Huttenhower, C, Jackson, MA, Janssen, S, Jeste, DV, Jiang, L, Kelley, ST, Knights, D, Kosciolek, T, Ladau, J, Leach, J, Marotz, C, Meleshko, D, Melnik, AV, Metcalf, JL, Mohimani, H, Montassier, E, Navas-Molina, J, Nguyen, TT, Peddada, S, Pevzner, P, Pollard, KS, Rahnavard, G, Robbins-Pianka, A, Sangwan, N, Shorenstein, J, Smarr, L, Song, SJ, Spector, T, Swafford, AD, Thackray, VG, Thompson, LR, Tripathi, A, Vázquez-Baeza, Y, Vrbanac, A, Wischmeyer, P, Wolfe, E, Zhu, Q and **Knight, R.** 2018. American Gut: An Open Platform for Citizen Science Microbiome Research. *mSystems*, 3(3). DOI: https://doi.org/10.1128/mSystems.00031-18

Milgram, S. 1974. *Obedience to authority: An experimental view*. New York, Harper & Row.

Nafus, D and **Sherman, J.** 2014. Big data, big questions|This one does not go up to 11: The quantified self movement as an alternative big data practice. *International Journal of Communication*, 8: 1784–1794.

National Center for Biotechnology Information, U.S. National Library of Medicine. n.d. GenBank Overview. Available at: https://www.ncbi.nlm.nih.gov/genbank/ [Last accessed 15 October 2018].

National Institutes of Health. 2018. Inclusion of women and minorities as participants in research involving human subjects – policy implementation page. Available at: https://grants.nih.gov/grants/funding/women_min/women_min.htm [Last accessed 15 October 2018].

Nest Labs. 2018. Privacy statement for Nest products and services. Available at: https://nest.com/legal/privacy-statement-for-nest-products-and-services/ [Last accessed 15 October 2018].

Open Humans Foundation. n.d.a About Open Humans. [Online]. Available at: https://www.openhumans.org/about/ [Last accessed 15 October 2018].

Open Humans Foundation. n.d.b The Personal Genome Project. Available at: https://www.personalgenomes.org/us [Last accessed 15 October 2018].

Quantified Self. 2018. Camille Nebeker: Informed Consent, Self-Consent. Available at: https://medium.com/quantified-self-public-health/camille-nebeker-informed-consent-self-consent-db63b842276e [Last accessed 15 October 2018].

Quantified Self Labs. 2015. Quantified Self: Self Knowledge Through Numbers. Available at: http://quantifiedself.com/ [Last accessed 15 October 2018].

Rolland, B, Lee, CP and **Potter, JD.** 2017. Greater than the sum of its parts: A qualitative study of the role of the Coordinating Center in facilitating Coordinated Collaborative Science. *Journal of Research Administration*, 48(1): 65–85.

Rothstein, MA, Wilbanks, JT and **Brothers, KB.** 2015. Citizen science on your smartphone: An ELSI research agenda. *The Journal of Law, Medicine & Ethics*, 43(4): 897–903. DOI: https://doi.org/10.1111/Fjlme.12327

Shirk, JL, Ballard, HL, Wilderman, CC, Phillips, T, Wiggins, A, Jordan, R, McCallie, E, Minarchek, M, Lewenstein, BV, Krasny, ME and **Bonney, R.** 2012. Public participation in scientific research: A framework for deliberate design. *Ecology and Society*, 17(2): 29. DOI: https://doi.org/10.5751/ES-04705-170229

Smith, A. 2013. Smartphone ownership - 2013 update. Washington, DC: Pew Research Center. Available at: http://www.pewinternet.org/2013/06/05/smartphone-ownership-2013/ [Last accessed 15 October 2018].

Swan, M. 2012. Crowdsourced health research studies: An important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research*, 14(2): e46. DOI: https://doi.org/10.2196/Fjmir.1988

U.S. Department of Health and Human Services. n.d. Federal policy for the protection of human subjects ('Common Rule'). Available at: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html [Last accessed 15 October 2018].

Vayena, E, Mastroianni, A and **Kahn, J.** 2012. Ethical issues in health research with novel online sources. *American Journal of Public Health*, 102(12): 2225–30. DOI: https://doi.org/10.2105/AJPH.2012.300813

Vayena, E, Salathe, M, Madoff, LC and **Brownstein, JS.** 2015. Ethical challenges of big data in public health. *PLoS Computaional Biology*, 11(2): e1003904. DOI: https://doi.org/10.1371/journal.pcbi.1003904

Vayena, E and **Tasioulas, J.** 2013. Adapting standards: Ethical oversight of participant-led health research. *PLoS Medicine*, 10(3): e1001402. DOI: https://doi.org/10.1371/journal.pmed.1001402

Weibel, N, Desai, P, Saul, L, Gupta, A and **Little, S.** 2017. HIV risk on twitter: The ethical dimension of social media evidence-based prevention for vulnerable populations. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 1775–1784. DOI: https://doi.org/10.24251/HICSS.2017.216

Wikilife Foundation. n.d. Datadonors|Bye bye! Available at: http://datadonors.org/ [Last accessed 15 October 2018].

Wolf, G. 2009. Know thyself: Tracking every face of life, from sleep to mood to pain, 24/7/365. *Wired*, 22 June [online access at: https://www.wired.com/2009/06/lbnp-knowthyself/ last accessed 15 October 2018].

Zimbardo, PG. 2018. *Stanford Prison Experiment*. Available at: http://www.prisonexp.org/ [Last accessed 15 October 2018].

How to cite this article: Bietz, M, Patrick, K and Bloss, C. 2019. Data Donation as a Model for Citizen Science Health Research. *Citizen Science: Theory and Practice*, 4(1): 6, pp. 1–11. DOI: https://doi.org/10.5334/cstp.178

Submitted: 18 May 2018 **Accepted:** 22 September 2018 **Published:** 08 March 2019

]u[*Citizen Science: Theory and Practice* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS