First, design for data sharing

John Wilbanks & Stephen H Friend

To upend current barriers to sharing clinical data and insights, we need a framework that not only accounts for choices made by trial participants but also qualifies researchers wishing to access and analyze the data.

This March, Sage Bionetworks (Seattle) began sharing curated data collected from >9,000 participants of mPower, a smartphone-enabled health research study for Parkinson's disease[1]. The mPower study is notable as one of the first observational assessments of human health to rapidly achieve scale as a result of its design and execution purely through a smartphone interface[2]. To support this unique study design, we developed a novel electronic informed consent process that includes participant-determined data-sharing preferences. It is through these preferences that the new data—including self-reported outcomes and quantitative sensor data—are shared broadly for secondary analysis. Our hope is that by sharing these data immediately, prior even to our own complete analysis, we will shorten the time to harnessing any utility that this study's data may hold to improve the condition of patients who suffer from this disease.

Turbulent times for data sharing

Our release of mPower comes at a turbulent time in data sharing. The power of data for secondary research is top of mind for many these days. Vice President Joe Biden, in heading President Barack Obama's ambitious cancer 'moonshot', describes data sharing as second only to funding to the success of the effort[3]. However, this powerful support for data sharing stands in opposition to the opinions of many within the research establishment. To wit, the august *New England Journal of Medicine* (*NEJM*)'s recent editorial suggesting that those who wish to reuse clinical trial data without the direct participation and approval of the original study team are "research parasites"[4]. In the wake of colliding perspectives on data sharing, we must not lose sight of the scientific and societal ends served by such efforts.

It is important to acknowledge that meaningful data sharing is a nontrivial process that can require substantial investment to ensure that data are shared with sufficient context to guide data users. When data analysis is narrowly targeted to answer a specific and straightforward question—as with many clinical trials—this added effort might not result in improved insights. However, many areas of science, such as genomics, astronomy and high-energy physics, have moved to data collection methods in which large amounts of raw data are potentially of relevance to a wide variety of research questions, but the methodology of moving from raw data to interpretation is itself a subject of active research.

It is our view that the emerging area of mobile health is another such area, and that data sharing has powerful potential to accelerate discovery. Rapid sharing of data from a large-scale observational study, such as mPower, provides a mechanism to distribute the task of developing appropriate analytical methods and identifying the approaches that maximize the utility of this new type of data. As with any new technology, a major obstacle to extracting clinical utility will be the development of useful analytical approaches. By facilitating the rapid and widespread distribution of mobile health data, we hope to entice a community of researchers to evaluate the applicability of a wide range of analytical approaches to achieve meaning from this emerging type of data.

Additionally, as researchers, we have an ethical obligation to participants to maximize the scientific value of their data donation. Our engagement with research participants should be as co-equals in the research ecosystem. Meaningful engagement with participants includes soliciting and honoring participant preferences for the distribution of their donation. Our experience suggests that participants who give their time and their sensitive personal information to researchers often assume that their data will be distributed widely to the full research community, not 'owned' as an asset to extract value from, solely by the researchers who happened to collect it. It is precisely to enable a new class of medical researchers that we at Sage Bionetworks offer participants the choice as to whether or not to share their own study data. In our view, those who would reuse study data are more commonly known as data scientists than parasites, and their reanalysis is to be welcomed.

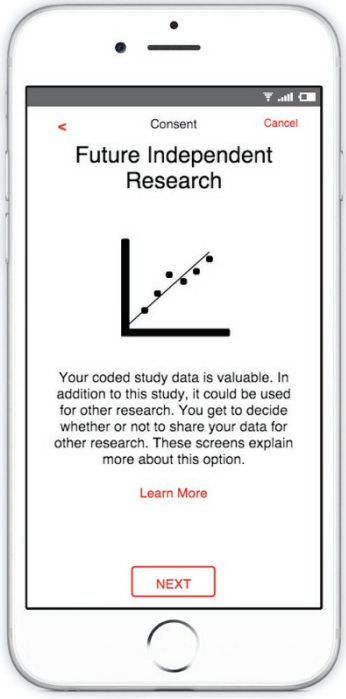

Figure 1 Image of toggle screen in mPower app.

Qualifying users and empowering participants

To address this misconception head-on, as part of the mPower informed consent, each

John Wilbanks and Stephen H. Friend are at Sage Bionetworks, Seattle, Washington, USA.
e-mail: friend@sagebase.org

research participant gets to decide whether or not to allow "qualified researchers worldwide" to access his or her coded study data. Additionally, participants can easily and independently change their data-sharing setting at any time during the study by means of the app's 'settings' screen. Given the choice, more than 75% (9,520 of 12,201) of consented and enrolled participants decided to share their data broadly (**Fig. 1**) (https://sagebionetworks.jira.com/wiki/display/BRIDGE/Bridge+Status#BridgeStatus-Parkinson.1).

We take the clear preference of study participants very seriously, and have been consulting with bioethicists so that we can draw on their decades of work to inform this nascent data-sharing framework. In the second half of 2015, we developed a process to qualify users for data access (https://www.synapse.org/#!Synapse:syn4993293/wiki/247860). We began by examining contemporary methods, most of which operate on a per-request evaluation model. Either the use, or the user, or most commonly both, goes through a Data Access Committee or other review mechanism, and matches data use restrictions on the original data gathering to the proposed use and/or user[5].

We chose to reject this model. First, we do not encode complex data use restrictions in our informed consent models, so we do not need complex review processes. All qualified users are welcome once they qualify. Second, the literature indicates that Data Access Committee mechanisms can encode conflicts of interest, leading to data withholding[6]. As a result, we felt a Data Access Committee would hinder widespread reuse of shared data from mPower, which is the entire point of our offering participants to share their data with qualified researchers worldwide.

We chose instead to prioritize the desires of our study participants, those choosing to share with "qualified researchers worldwide," in conceptualizing our data-sharing framework. We imposed few use restrictions: we excluded commercial resale, marketing uses and re-identification of data donors. We elected not to differentiate researchers based on the tax status of their employer. Although many sharing systems differentiate between academic and corporate use, we noted the extremely common practices of corporate-sponsored research at universities, or of corporate partnership with nonprofits in rare disease—making this distinction not only difficult to define, but capable of substantially blocking research. We also evaluated and rejected restrictions on what diseases or protocols may be investigated—such differentiation can block investigations of adjacent diseases and discovery of underlying integrative biology.

But this broad interpretation of data sharing left us wondering if we had properly balanced the beneficence expected from widespread data reuse with the ethical obligations to data donors. We therefore looked to transactional programs based on trust, such as the 'Global Entry' traveler program in the United States, as a method that could mitigate many predictable risks of broad data sharing. Although anyone is eligible to apply for such programs, only those who pass an initial screening and background check get in. Thus, the next phase of development was to decide on what screens and checks should apply to researchers aspiring to become qualified researchers and gain access to donated data.

Our goals in designing the data-sharing procedure were the following:

- balance the expected protection of participants' privacy with their desire for optimal data use and reuse;
- emphasize transparency, so that anyone can know how data are being used and by whom;
- describe and cultivate a clear set of behavioral norms for working with participant-donated data sets;
- assess data requester's knowledge of research ethics and appropriate conduct for accessing, using and managing participant-donated data;
- emphasize return of information, data and results to participants and the research community.

To achieve these goals, we require all data requestors become "qualified researchers" by completing the following steps:

- demonstrate their awareness and understanding of the data-sharing framework and applied ethics through a short, 18-question examination;
- validate their identity to Sage Bionetworks through a variety of approved methods, such as an academic letter from a signing official, a notarized letter attesting to identity or a copy of a professional license;
- make a public statement of intended data use, which we can in turn feed back to participants in the spirit of engagement and transparency;
- explicitly agree to a 'contract' of data sharing, including the following: (i) downloading, initialing, signing, scanning and uploading a researcher oath to adhere to a code of behavior; (ii) complying with any data-specific conditions of use.

Taken together, we believe this process creates a set of transactional requirements that will not block any dedicated researcher—but dramatically increases the chances that researchers who do become qualified act in an ethical manner with the donated data.

Challenges and next steps

Our next task is to develop robust, participant-centered enforcement and dispute resolution processes so that those who make the gift of data have a voice in deciding when data use has gone awry. Multiple kinds of potential disputes exist; for example, criminal use of data, allegations of breach of the contract of data sharing, and unethical activity, and each has different kinds of resolution mechanisms available.

For allegations of criminal behavior, dispute resolution is perhaps simplest: we will refer the matter to the relevant law enforcement agency. For breach of contract, we have an array of options, ranging from reprimands to banning the user permanently from our platforms to pursuing tort cases in the courts. For allegations of unethical behavior, we are looking into options that directly engage data donors, such as a participant-led board that has explicit authority to review and decide on specific allegations. We are also exploring how relationships with professional societies and patient advocacy groups might help reinforce an ecosystem of ethical data reuse. The approaches that we can control completely alone are few; we will have to work dynamically within existing structures as well as create new structures for evaluation and reaction as practice informs theory. This is one of many reasons to work as hard up front to make improper data use though inattention or malice as unlikely as possible.

The above set of 'probable' violations is what we have explored thus far; we anticipate there will be others. The novelty of our approach will land squarely into a clinical practice reality in which researchers often consume data cavalierly—laptops left on buses, USB drives left on laboratory benches, data shared *sub rosa* through DropBox—and we do not wish to be caught unaware. Our goals are to minimize these behaviors through awareness-raising in the qualification process, but we must be ready to decisively act on evidence of misuse of patient-donated data. This is part of the deal: to access a new kind of data, researchers must act in a new way.

Although we have worked diligently to create a robust data-sharing framework, we recognize the limitations of our ability to predict the consequences of this kind of sharing, whether for good or bad. Thus, we are releasing data just from one of our mobile health studies—the mPower Parkinson's survey—to

begin. The mPower data set is derived from quantitative sensor measurements (e.g., gyroscope, accelerometer, touchscreen, microphone) collected during specified 'active tasks' in our mobile study application, as well as participant responses to clinical surveys, such as UPDRS[7]. The mPower study is purely observational; all data are coded and scrubbed of directly identifiable participant information to further mitigate the risk of data sharing.

Inertia in the clinical investigator population is another challenge to our data-sharing framework, one that was perhaps revealed by the *NEJM* editorial. Indeed, we often hear that investigators are accustomed to exclusive control of 'their' data sets and that our desire to give this right to enrolled participants is highly irregular. To offset this 'irregularity', thanks to a grant from the Robert Wood Johnson Foundation (Princeton, NJ, USA), we are able to offer a substantial incentive to investigators wishing to adopt our approach: the ability to run mobile clinical studies through our platform at costs an order of magnitude lower than they would be otherwise. In exchange, our tax is one of participant-centricity rather than profit.

Another challenge we foresee is that participant desire to share data may be contextual. It is easy to choose to share data broadly in the abstract; that choice is likely to change in unpredictable ways as data reuse becomes concrete. Some participants may think of researchers exclusively as academics and choose to turn off broad sharing when they recognize the breadth of the community of access. In contrast, others may be encouraged by research progressing at a pharmaceutical company and decide to turn on their broad sharing option. This ability to toggle data-sharing options cannot be taken lightly. Data already shared and included in secondary research cannot be called back. We are thus developing information loops to support decision making around the sharing setting inside our studies and to increase the awareness of participants about reuse and their ability to see who is using their data and for what purposes.

Perhaps the most commonly raised challenge by outside researchers and clinical study sponsors is the uncertainty of securing ethical and regulatory approval when adopting this new approach. Our own experience with ethical review, both for our own protocols and for the protocols of partners using our technology platforms, has been uniformly positive. We spent much of 2014 and 2015 in active engagement with the institutional review board (IRB) and bioethics communities, and found groups eager to engage in solutions to long-standing problems with informed consent and data reuse. Both the IRB and bioethics communities bring essential diversity and insight to issues of data reuse and informed consent—but far too often they are engaged late, if at all, and labeled as 'blockers' of the will of investigators or patients. This is a situation where input from the professional ethics community, and engagement, integration and ethnography have helped enormously.

Finally, it is important to acknowledge that our experience comes from running a relatively novel kind of clinical study, one based on mobile devices and patient-generated data. But there is nothing in the methodology that means our proposal should be limited to clinical studies using smart phones; quite the opposite. As we have seen in the past, any study can put the participant at the center of decisions around data sharing—it is a matter of will, not a matter of technology[8].

Sage Bionetworks' qualified researcher process doesn't solve all issues of clinical trial data sharing. Nor does it provide a single answer on how to center the participant in research. This is a beginning, not an end. There remain essential issues of technology security, privacy, governance and study design. But clinical studies should be contemplating far more comprehensive data reuse policies, or else face a growing sense that the parasite-host relationship might be one in which the investigators themselves are on the wrong side.

Published online 3 March 2016; doi:10.1038/nbt.3516

COMPETING INTERESTS STATEMENT
The author declares competing financial interests: details accompany the online version of the paper (doi:10.1038/nbt.3516).

1. Bot, B.M. *et al.* The mPower study, Parkinson's disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016). doi:10.1038/sdata.2016.11
2. Trister, A., Dorsey, E.R. & Friend, S.H. Smartphones as new tools in the management and understanding of Parkinson's disease. *npj Parkinson's Disease* **2**, 16006 (2016); doi:10.1038/npjparkd.2016.6; published online 3 March 2016.
3. Biden, J. Jr. Inspiring a new generation to defy the bounds of innovation: a moonshot to cure cancer. https://medium.com/@VPOTUS/inspiring-a-new-generation-to-defy-the-bounds-of-innovation-a-moonshot-to-cure-cancer-fbdf71d01c2e#.gx72sbluo (12 January 2016).
4. Longo, D.L. & Drazen, J.M. *N. Engl. J. Med.* **374**, 276–277 (2016).
5. Shabani, M., Knoppers, B.M. & Borry, P. *EMBO Mol. Med.* **7**, 507–509 (2015).
6. Shabani, M., Dyke, S.O.M., Joly, Y. & Borry, P. *PLoS Biol.* **13**, e1002339 (2015).
7. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. *Mov. Disord.* **18**, 738–750 (2003).
8. Forms, C. Framingham Heart Study. Visited 8 February 2016. https://www.framinghamheartstudy.org/researchers/concent-forms.php