Person-Centered Medical and Healthcare Studies

Ross Gore
Virginia Modeling, Analysis and Simulation Center
Suffolk, VA
rgore@odu.edu

Manasi Sheth-Chandra
Center for Global Health
Norfolk, VA
msheth@odu.edu

ABSTRACT

Person-Centered Studies (PCS) are a new approach for the design and analysis of medical and healthcare research with human participants. The PCS approach is based on the idea that data can be privately maintained by participants and never revealed to researchers, while still enabling medical and healthcare statistical models to be fit and research hypotheses to be tested. PCS rests on the assumption that data should belong to, be controlled by, and remain in the possession of participants in studies. Since data have value, individuals can accumulate personal wealth by participating in medical and healthcare science. The key observation behind the PCS approach is that medical and healthcare statistical models can be fit by sending an objective function and vector of parameters to each participants' smartphone, where the likelihood of that participants' data is calculated locally. Only the likelihood value is returned to the central optimizer. The optimizer aggregates likelihood values from all participants and chooses a new vector of parameters until the model converges. This transformative workflow relies on two modeling components: (1) a *medical-data dropbox* (MDB) for patients to maintain possession of their individual data and not reveal it and (2) a method for *scattered likelihood estimation* (ScaLE) so that each participant's smartphone calculates the likelihood of their own data and passes only the likelihood value back to a centralized optimizer. The PCS approach solves or simplifies many current problems that plague medical and healthcare research. A PCS study provides: (1) significantly greater privacy for participants, (2) lower cost for the researcher and funding institute, (3) a larger base of participants and (4) faster determination of results.

ABOUT THE AUTHORS

Ross Gore holds a Doctorate of Philosophy and a Master's degree in Computer Science from the University of Virginia and a Bachelor's degree in Computer Science from the University of Richmond. Dr. Gore has ten years of research experience in problems that lie at the intersection of computer science and modeling and simulation. His work has yielded authorships on more than 20 conference and journal publications and has been recognized by the ARCS (Achievement Rewards for College Scientists) Society as an impactful and novel research avenue.

Manasi Sheth-Chandra holds a PhD in Statistics as well as a M.S. in Statistics from Old Dominion University. She also holds a M.S. in Mathematics and a B.S. in Mathematics as well as Biochemistry from Loyola University of Chicago. She has worked on various projects at Booz Allen Hamilton, supporting clients at NASA Langley as well as Navy Marine Corps Public Health Center. She has helped to apply moderately advanced statistical methods, assisted in producing analyses plans, statistical reports, and mentored departmental members. Her teaching experience includes introductory as well as intermediate college level mathematics and statistics.

Person-Centered Medical and Healthcare Studies

Ross Gore
Virginia Modeling, Analysis and Simulation Center
Norfolk, VA
rgore@odu.edu

Manasi Sheth-Chandra
Center for Global Health
Norfolk, VA
msheth@odu.edu

INTRODUCTION

The traditional research paradigm for medical and healthcare sciences is that data are collected into a centralized repository. After data collection, this central data repository is used to fit candidate statistical models via estimates of model parameters. We propose Person-Centered Studies (PCS), a transformative alternative to the traditional research paradigm. In PCS, data remain where they were originally collected, on each participant's smartphone, and remain private. Candidate statistical models are fit by using a secure internet connection to send a vector of model parameters to each person's smartphone. Each smartphone then calculates the likelihood of the data and only this single likelihood value is returned to the research lab. An optimizer at the research lab aggregates likelihood values from all the scattered participant devices and chooses a new set of parameters and the cycle begins again. This process repeats until the models converge to a maximally likely set of parameter values.

PCS relies on two components: (1) a medical-data dropbox (MDB) for patients to maintain possession of their individual data and not reveal it and (2) a method for scattered likelihood estimation (ScaLE) so that each participant's smartphone can calculate the likelihood of their own data and pass only the likelihood value back to a centralized optimizer. When using the PCS approach data can be collected at the same time that models are optimized. This means that when sufficient data are collected to reach a pre-selected statistical power, the study can automatically terminate or switch to a cross-validation regime. This means that individual-level variables, repeated measurements, and time series are automatically linked and model parameters are estimated at the individual- level first and only then does aggregation happen. The result of these features is a new approach to how health science research is designed, conducted, and analyzed by creating a system where data are collected, statistical analyses are conducted, and scientific hypotheses tested while: (1) individuals remain in their normal living environment, (2) individuals maintain possession and ownership of their data and (3) individuals do not ever reveal their data.

To elucidate this impact, we provide an overview of traditional experimental design and statistical estimation. Next, we describe the medical-data dropbox (MDB) portion of PCS for smartphones. This is followed by a description of a method of scattered likelihood estimation (ScaLE) that PCS requires. Then we demonstrate how these two components can be combined into a workflow (PCS) for a transformative shift in medical and healthcare research with human participants. The demonstration highlights the challenges that must be completed to realize PCS. Finally, we summarize the benefits of PCS to medical and healthcare science, to individual researchers, and to research participants.

TRADITIONAL EXPERIMENTAL DESIGN

Traditional research practices in medical and healthcare science have a common sequence of events depicted in Figure 1. First a hypothesis is generated. Next, an experiment is designed that tests the hypothesis. Participants are recruited and those that consent are enrolled to participate in the experimental protocol. Data are collected and centralized into a data repository that typically resides on a computer in a locked room in a research laboratory. These data are generally considered to belong to the research group that performed the experiment or collected the observational data.

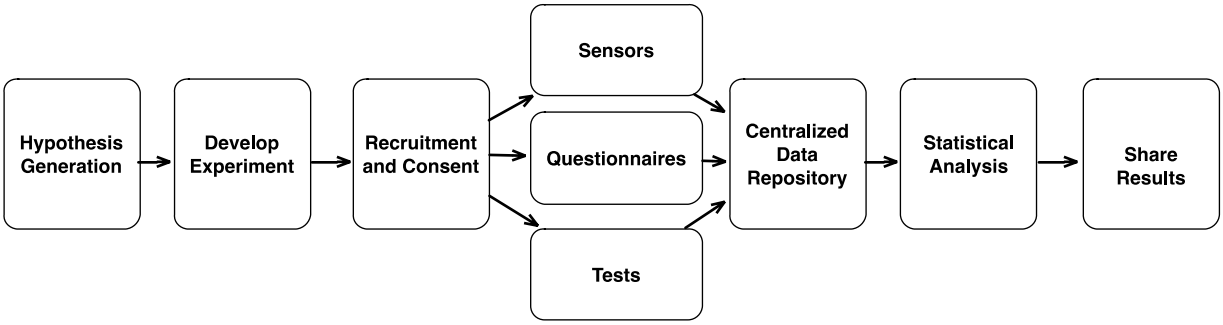

Figure 1. Traditional Experimental Design

After the data collection phase, the centralized data are analyzed to estimate parameters of candidate statistical models and test the original hypotheses. One of the most commonly used procedures to estimate parameters is full information maximum likelihood (FIML). Within FIML, the data collected for the candidate models are stored in a matrix. Then, starting values for the model parameters are selected. Given the starting values of the parameters, the model implies an expected covariance and means structure for the data. For each row of the data matrix the likelihood of the data is calculated given the model parameters. These likelihoods are logged, summed and a test is performed to see if this summed log likelihood is at a maximum. If the summed log likelihood is not at a maximum, a new set of parameters are chosen and new likelihoods are calculated. If the summed log likelihood is at a maximum, the software returns the summed log likelihood and the current parameter estimates. The process of FIML estimation is shown in Figure 2.

Once the parameters estimates are complete and the original hypotheses are tested, the results are disseminated through journal articles and/or conference talks and posters. Finally, the experimental results are replicated and/or new hypotheses are generated.

Despite the widespread use of this approach to medical and healthcare research with human subjects, there are at least four potential problems:

1. It is unclear who owns the data. Is it the research group who collected the data, the agency that funded the research, or do the participants retain ownership of their own data?
2. Barriers to data sharing include protecting the confidentiality of the participants and a potentially long lag time during which the research group has the sole right of publication using the data.
3. During the process of designing a new experiment, power analyses (when they are conducted) are encouraged to be conservative, since one does not know the effect size in advance.
4. It is difficult, if not impossible, to perform longitudinal linking between different data repositories while maintaining participant confidentiality. Thus longitudinal studies tend to be isolated from the benefits of data sharing.

In the next section, we present a person-centered approach for medical and healthcare research using human participants. Our approach allows participants to maintain possession, ownership, and control of their own data while still allowing statistical tests of scientific hypotheses even if individuals never divulge their data. Following this approach to its logical conclusion leads to efficiencies and simplifications to many of the problems enumerated above.

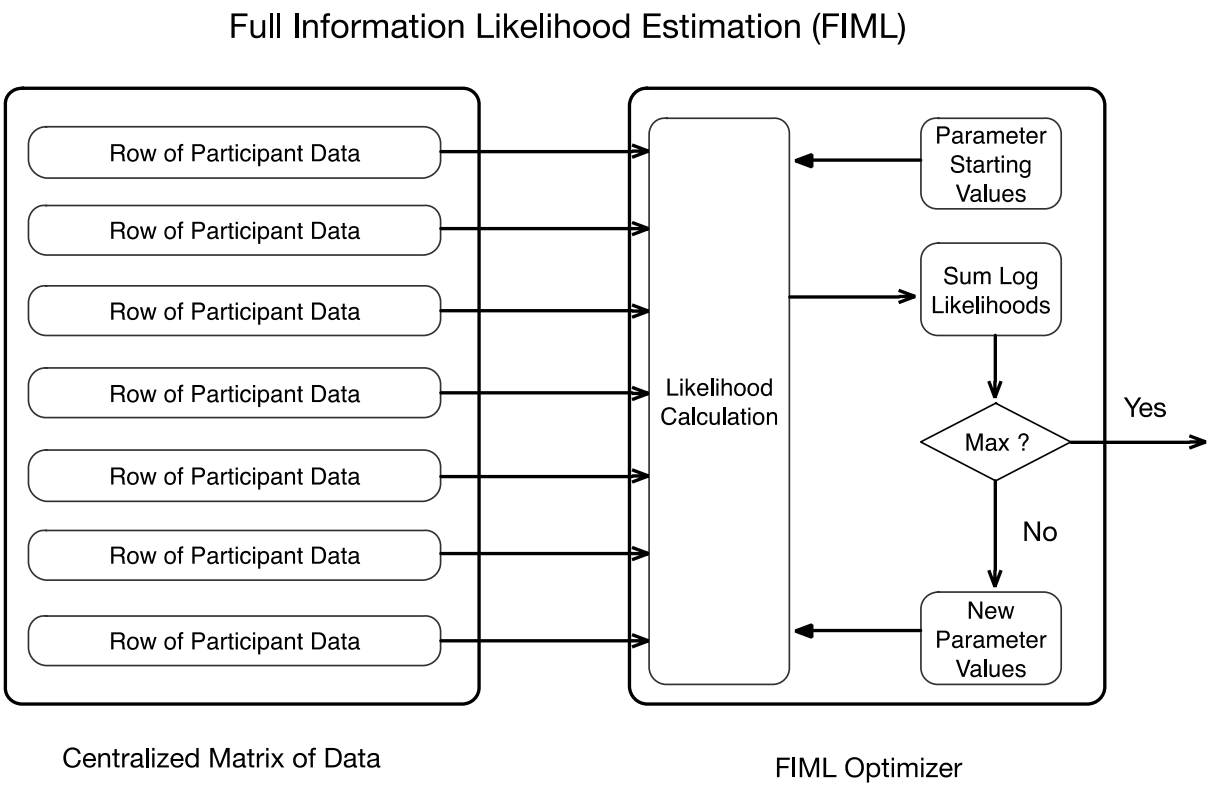

Figure 2. Full Information Likelihood Estimation.

PERSON-CENTERED STUDIES (PCS)

Person-Centered Studies (PCS) address the limitations of traditional experimental design for medical and healthcare research with human participants. There are two parts to PCS: (1) that individuals maintain possession of their data and need not reveal their data; and (2) that scientific hypotheses can be tested against participants' private data by scattering the statistical likelihood estimation so that each participant's smart phone calculates the likelihood of their own data and passes only the likelihood value back to a centralized optimizer. The technologies which implement these two parts are: (1) the medical-data dropbox (MDB) and the method for *scattered likelihood estimation (ScaLE)*. First, we present each of these pieces of technology independently. Then we describe how their interactions define the PCS research approach.

Person-Centered Medical DropBox (MDB)

Data collection using participants' own devices such as personal computers, smartphones, and tablets has become increasingly common over the past 20 years (Dufau et al., 2011; Miller, 2012). Methods associated with this type of data collection are often longitudinal in nature and go by names such as ecological momentary assessment (Shiffman, Stone and Hufford, 2008; Shiffman and Stone, 1994), experience sampling (Hektner, Schmidt and Csikszentmihalyi, 2007; Larson and Csikszentmihalyi, 1983; Csikszentmihalyi and Larson, 1987), and intensive longitudinal designs (Walls and Schafer, 2006). As technology has advanced, data from these experiments has been collected by email (Lazer et al., 2009), web browsers (Greenwald and Nosek, 2001), social networking applications (Anderson et al., 2012) and most recently by smartphone (Miller, 2012; Benocci et al., 2010). Although the technology for data collection has improved remarkably, the basic scientific workflow has remained the same in that data are first collected into some central repository and only then are they analyzed.

In this paper we present a medical-data dropbox (MDB) that runs on participants' smartphones and serves to communicate between the research group sponsoring a study and an experiment app disseminated by the research

group and downloaded by the participant. The experiment app features methods for data acquisition, while the MDB stores all the participant's medical data with encryption and only allow access through a very specific set of communication protocols. Specifically, the MDB:

1. Acts as an intermediary that would allow research groups to advertise for participants. The MDB owner could browse for studies in which she or he wanted to participate.
2. Manages consent forms in a standardized way so that the potential participant could opt-in or opt-out of an experiment at any time.
3. Provides a data socket that would communicate with the experiment app that presented the experimental stimulus, questionnaire, test, or other data acquisition method.
4. Encrypts and maintain all participant data.

Many computer applications have been proposed to store health data for large scale applications [6, 2] and on personal devices like smartphones (Mandl et al., 2007; Sunyaev et al., 2010). However, these applications are designed to simply organize and protect personal health data and are not designed to be able to participate in statistical model fitting. Next, we present a method that allows smartphones scattered across large-scale networks to be able to participate in scientific experiments without revealing the data stored by MDB on the smartphone.

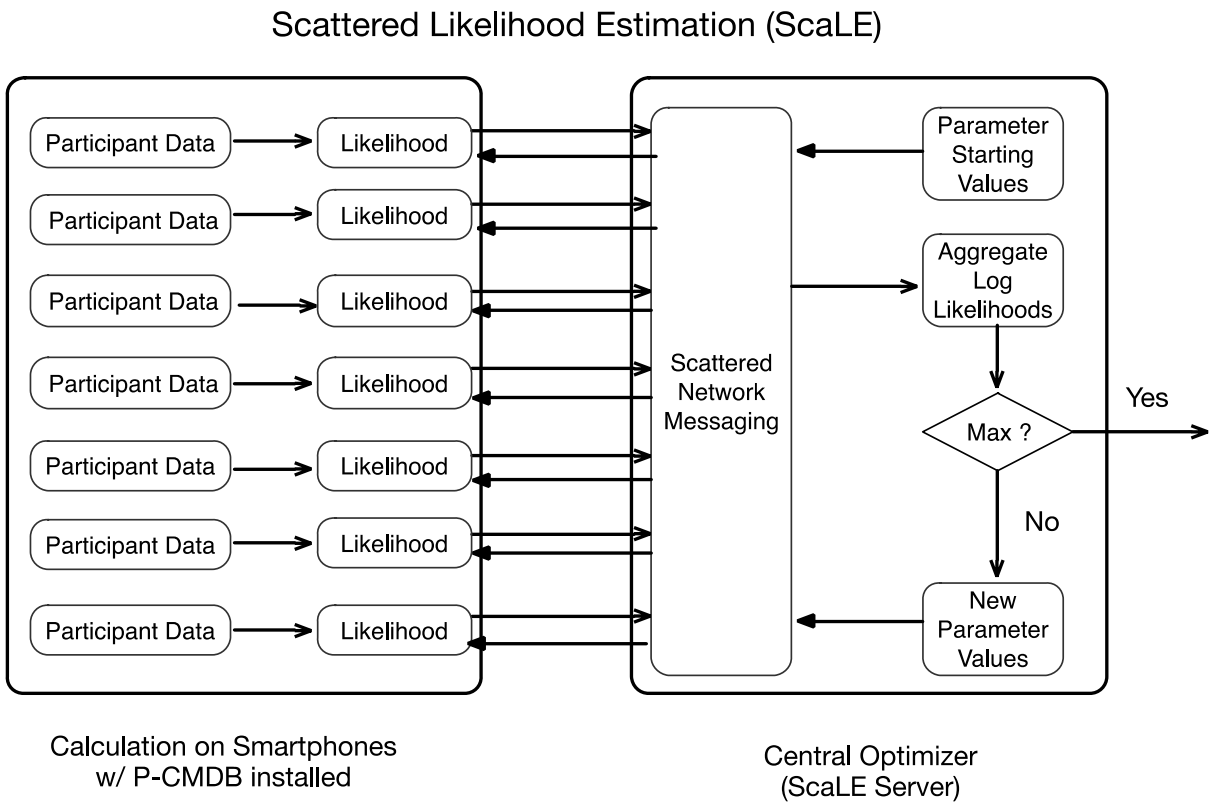

Figure 3. Scattered Likelihood Estimation.

Scattered Likelihood Estimation (ScaLE)

Scattered likelihood estimation (ScaLE) refers to a method for estimating parameters for statistical models that is very similar to the FIML method presented earlier. Figure 3 shows the ScaLE estimation process. As in FIML, a model and starting values for the parameters are selected by the researcher. The model and parameter values imply a covariance and means structure such that the likelihood of a single participant's data can be calculated. The model and parameters are scattered from a central optimization server (the ScaLE server) to all of the smartphones that

have consented to allow the analysis. Each participant's smartphone uses its MDB to calculate the likelihood of the data stored on the device. The MDB then sends only the likelihood number back to the ScaLE server. The ScaLE server then aggregates the log likelihoods from the responding smartphones with MDB installed. After aggregation the ScaLE server either decides it is at a log likelihood maximum or adjusts parameters and redistributes the new parameters to the smartphones with MDB installed. As a demonstration of the feasibility of this approach, we simulated longitudinal data from a latent growth curve model (McArdle and Epstein, 1987) and created a sampling design where (a) individuals opt into a study at a fixed rate up to a maximum enrollment, (b) individuals fill in a repeated measures questionnaire with a fixed amount of time between occasions of measurement, and (c) individuals have an 80% probability of having their smartphone on at any particular moment. We fit these simulated data using a distributed likelihood approach and the results are shown in Figure 4. Our results show that even though the sample size is not strictly increasing the model converges in less than 500 total iterations.

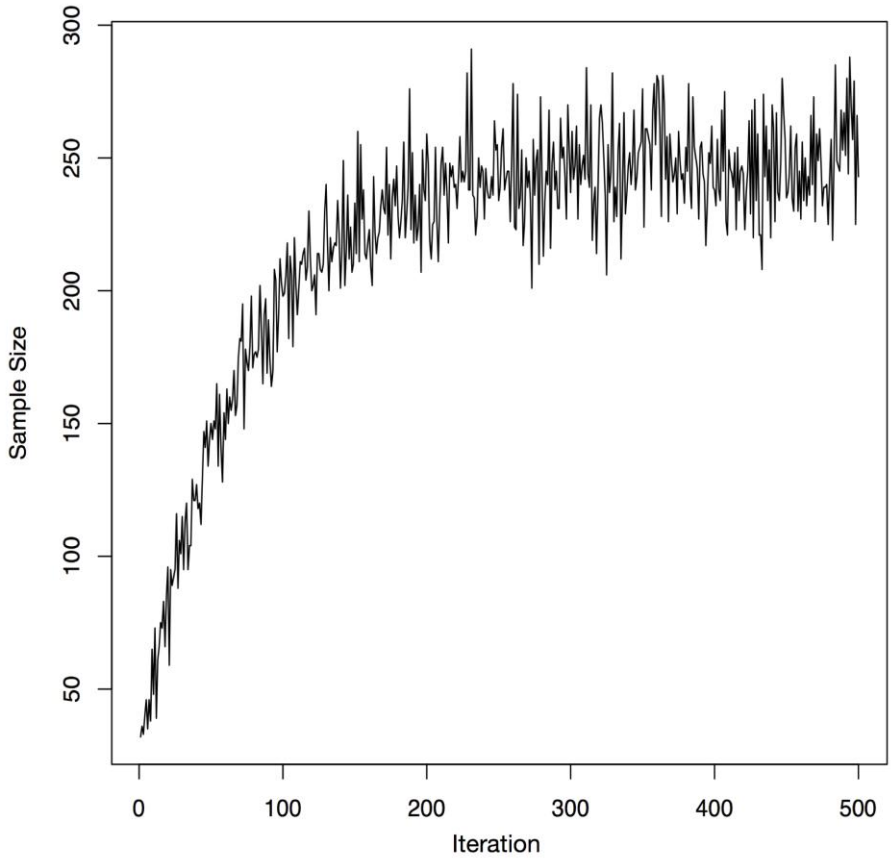

Figure 4. Sample size versus time in a ScaLE study with realistic participation parameters.

The similarity between FIML and ScaLE is evident, but the two algorithms are not identical. There are two major differences: (1) traditional FIML requires that all participants reveal their data to the researcher whereas ScaLE only requires that participants reveal the likelihood of their data given a model and a vector of parameters and (2) traditional FIML performs all of its calculations and data manipulations centrally whereas ScaLE performs the greatest part of its calculations in parallel on the participants' smartphones.

Several other differences are less apparent, but equally as important. ScaLE does not require that data collection be finalized prior to the initiation of model estimation. Statistical models can begin to be tested as soon as participants have consented and begun the experiment. Of course, model parameters will be unreliable with fewer participants. However, as more individuals consent and participate in the experiment, the sample size contributing to each likelihood calculation will grow, and the model parameters will begin to become more and more stable. This means

that a researcher can preselect a required statistical power or parameter precision for the analysis and the ScaLE analysis can automatically terminate or initiate a replication when the experiment reaches that criterion.

Also, since traditional FIML requires centralized data, longitudinal studies must link variables over time by some sort of identifying information so that data belonging to a single individual can be grouped. On the other hand, in a ScaLE analysis, all the data belonging to a participant and only the data belonging to a participant is available within the MDB software. As long as the participant consents to longitudinal linking of her or his data, the likelihood of the longitudinal data can be automatically calculated, even if some data resulted from a previous study. Thus, data sharing between experiments is up to the individual participant. If the participant consents to data sharing, data are automatically available and linked by participant without the researcher needing to become involved in complex data sharing agreements. This is a consequence of the data being owned, controlled and possessed by the participant.

PCS: Combining MDB and ScaLE

The PCS approach transforms how medical and healthcare research with human participants is conducted by combining MDB and ScaLE. In order to demonstrate how this combination works in practice, we describe the workflow of a PCS study. The study demonstrates the benefits of the PCS approach and illustrates the inner workings of the components in a PCS study.

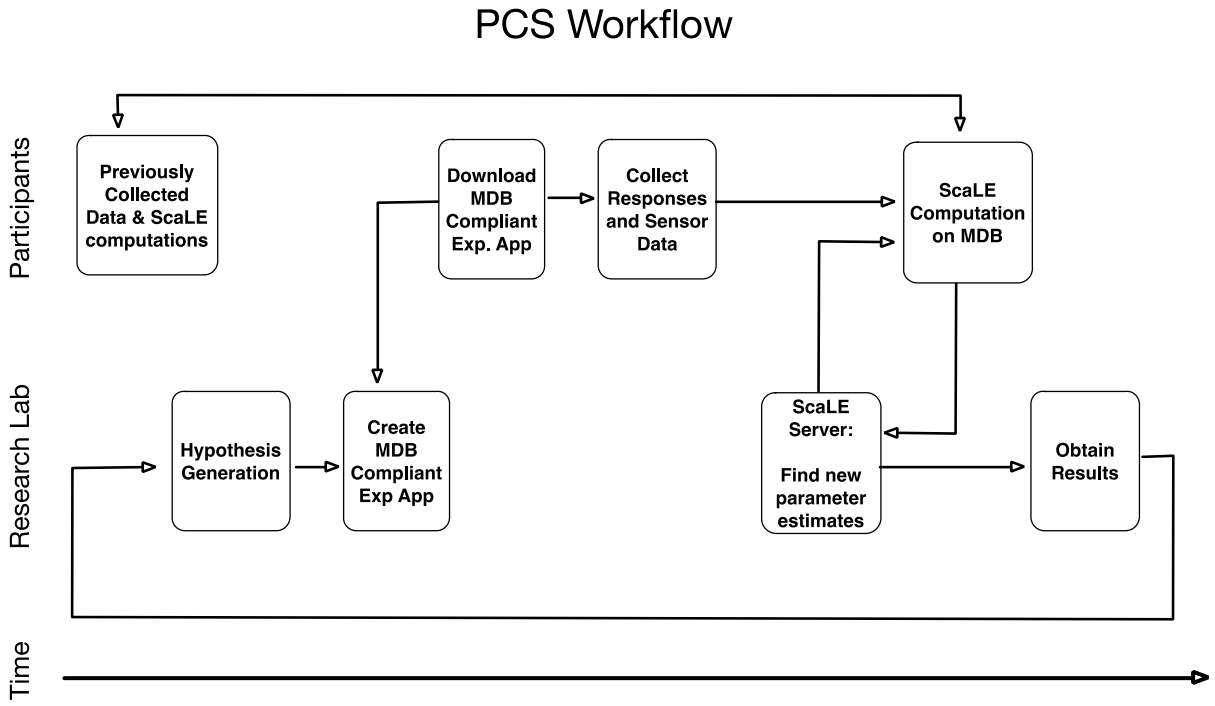

Figure 5. The workflow of a PCS study.

Figure 5 depicts the workflow of a PCS study. A research group generates a hypothesis and a MDB compliant experiment app featuring: (1) a self-report questionnaire instrument, (2) a tie-in to accelerometer sensor data and (3) a candidate statistical model is created at the research laboratory. Since the app is MDB compliant it is advertised to all individuals who have MDB installed on their smartphone. Some individuals become participants by downloading the experiment app, consenting to the study, taking the questionnaire and accelerometer readings. Once participants consent into the study model fitting via ScaLE proceeds as follows:

1. The ScaLE server chooses starting values for all parameters and sends these to all current participants.
2. Each participants' MDB calculates the likelihood of the participant's data collected so far and sends that likelihood number back to the ScaLE server.
3. The ScaLE server aggregates all the likelihood numbers from the MDBs.
4. The ScaLE server chooses new parameters and repeats the process until convergence criteria are reached.

5. The experiment is finished after collecting just-sufficient data and the results are obtained. The results are then disseminated to generate new hypotheses for other research labs.

Note that when participants give consent for the use of previously-collected data, each new experiment starts optimization with a large set of data automatically shared from previously-run PCS experiments. Longitudinal data collection is thus automatically enabled and linked at the individual level at zero cost to the new project. Furthermore, participants may choose to collect personal data in their MDB on their smartphone without previously opting into a study (e.g., using wearable activity monitors, health monitors, GPS tracking, etc.) and then track their own personal trends. If these participants later download an experiment app, they can choose to allow access of these previously collected data in the new experiment.

DISCUSSION

A number of problems must be solved in order to realize PCS. Security and privacy must be excellent. It should be noted that encryption is not the same as privacy. Of course, data on the MDB and transmissions among MDBs and PCS servers will need to be encrypted. But, no encryption is totally secure, it is merely expensive and difficult to break.

However, the PCS approach focuses on privacy, which reduces the payoff to a potential attacker who manages to intercept and decrypt transmissions on the PCS network, or from malicious actors within a PCS study. While the data on a given MDB would still be susceptible to a determined decryption attack, our approach to data ownership improves data privacy by decentralizing participant data. The potential reward for decrypting a single MDB is much lower than the reward for a successful attack on a current centralized repository. Any given MDB is therefore a much less attractive target for identity thieves. While no system will ever be entirely secure or private, risks to privacy will be substantially mitigated by PCS.

Some PCS studies will require group membership information. For instance, an analysis of social networks or family relationships will require individual participants to be identified with a group. This group membership information is data and as such should be treated as belonging to the individual. A method must be implemented where an individual can opt–into a group membership. One solution would be for group members to choose a common pass phrase and give that to participants when they opt–in. Multigroup model membership could then be incorporated into the PCS experiment software uploaded to each group member's MDB. Once the group membership data are stored in the MDB, it is straightforward to calculate objective functions conditional on group membership. However, optimization will require aggregating objective function values conditional on group membership. In order for group membership to not be revealed to the central optimizer, one solution would allow peer–to–peer aggregation of objective functions prior to their being transmitted to the PCS optimizer.

A modification of current best practices in informed consent will need to be developed. The user interface for obtaining consent from a PCS participant will need to include a variety of options that are not contained in standard consent documents. For instance, longitudinal linking and sharing across experiments can be separate consent items. Risk management options can be included in consent documents. Some individuals may be willing to allow plots of selected raw data to be generated locally and transmitted to the research study. Others may wish to only reveal function values.

Data privacy requires that data not be disclosed. But also, data must not be lost. Mechanisms for secure data backup must be available for MDBs. Individuals must be given a choice of backup mechanisms. One reasonable choice would be available an encrypted backup onto a cloud facility (Bhadauria et al., 2011; Jansen and Grance, 2011). Some participants may wish to only maintain a private MDB backup in their homes. Hospitals and/or primary care physicians may choose to offer access to encrypted cloud–based MDB backup as part of their health care services. Exploring the extent to which these factors improve or degrade privacy and security in PCS remains future work.

CONCLUSION

The basic premise of the PCS approach is that participants' data remain the personal property of each individual, thereby transforming the economic model of large–scale research both public and private. We believe that this philosophic shift is as revolutionary as when ownership of private property becomes allowed in a formerly

command–driven economic system. We predict that a market–driven personal data economy will arise as individuals realize that their personal data are personal property and has accumulating worth directly related to the data's quality and scarcity. Unforeseen innovations will surely arise from this new market–driven personal data economy. We are confident that as the PCS approach becomes widespread, the pace of innovation and discovery in the behavioral, social and health sciences will be vastly accelerated while risks to individual privacy will be mitigated relative to current research practice.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of our colleagues at the Virginia Modeling, Analysis and Simulation Center.

REFERENCES

Anderson, B., Fagan, P., Woodnutt, T., & Chamorro-Premuzic, T. (2012). Facebook psychology: Popular questions answered by research. Psychology of Popular Media Culture, 1(1), 23.

Benocci, M., Tacconi, C., Farella, E., Benini, L., Chiari, L., & Vanzago, L. (2010). Accelerometer-based fall detection using optimized ZigBee data streaming. Microelectronics Journal, 41(11), 703-710.

Bhadauria, R., Chaki, R., Chaki, N., & Sanyal, S. (2011). A survey on security issues in cloud computing. arXiv preprint arXiv:1109.5388.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. The Journal of nervous and mental disease, 175(9), 526-536.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X. & Grainger, J. (2011). Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. PloS one, 6(9), e24974.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. Zeitschrift für Experimentelle Psychologie, 48(2), 85-93.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (Eds.). (2007). Experience sampling method: Measuring the quality of everyday life. Sage.

Jansen, W., & Grance, T. (2011). Guidelines on security and privacy in public cloud computing. NIST special publication, 800, 144.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. New Directions for Methodology of Social & Behavioral Science.

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 (New York, NY), 323(5915), 721.

Mandl, K. D., Simons, W. W., Crawford, W. C., & Abbett, J. M. (2007). Indivo: a personally controlled health record for health information exchange and communication. BMC medical informatics and decision making, 7(1), 25.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. Child development, 110-133.

Miller, G. (2012). The smartphone psychology manifesto. Perspectives on Psychological Science, 7(3), 221-237.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. Annu. Rev. Clin. Psychol., 4, 1-32.

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. Annals of Behavioral Medicine.

Sunyaev, A., Chornyi, D., Mauro, C., & Krcmar, H. (2010, January). Evaluation framework for personal health records: Microsoft HealthVault vs. Google Health. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (pp. 1-10). IEEE.

Walls, T. A., & Schafer, J. L. (Eds.). (2006). Models for intensive longitudinal data (pp. 3-37). New York:: Oxford University Press.