Privacy and Health Research in a Data-Driven World

An Exploratory Workshop

Sponsored by the Office for Human Research Protections (OHRP), Department of Health and Human Services (HHS)

September 19, 2019

Privacy and Health Research in a Data-Driven World

OHRP Exploratory Workshop: September 19, 2019

Welcome and Introduction

- Jerry Menikoff, M.D.; Director, Office for Human Research Protections (OHRP)
- Yvonne Lau, MBBS, MBHL, Ph.D.: Director, Division of Education and Development, OHRP

Dr. Menikoff, Director of the U.S. Department of Health and Human Services (HHS) Office for Human Research Protections (OHRP), welcomed everyone to the meeting and thanked OHRP staff from its Division of Education and Development (DED) for planning the workshop. He explained that OHRP's goal is to provide an open forum to explore critical issues. Today's world is one in which unimaginable amounts of data are used for research purposes, including data collected in connection with clinical care. Members of the public also generate large amounts of data. There is "tremendous potential to use these data to benefit everyone," including the opportunity to discover new insights into treatment and innovative ways to support health. However, there is no consensus about the best ways to collect, store, and share these data ethically. OHRP staff are here as listeners and to collaborate with colleagues to find the best ways to support ethical research.

Dr. Lau, Director of DED, also welcomed everyone to the workshop. She explained that DED's mission is to conduct public outreach and education. The workshop is on a timely topic and features a panel composed of diverse experts. It begins with an exploration of the landscape of big data research and the challenge of promoting public good while respecting privacy concerns. The second session provides an opportunity to hear how various entities are addressing these challenges. The workshop concludes by highlighting roles and responsibilities of different entities in identifying and responding to challenges. Each session features an hour of discussion involving all members of the panel.

Note: This report provides only highlights of speakers' presentations. For an in-depth look at what each speaker had to say, please see the presenters' slides.

Session I: Is Privacy a Casualty of Advancing Research?

- *Moderator:* Jodi Daniel, J.D., M.P.H.; Crowell & Moring LLP

Session I Introduction

- Jodi Daniel, J.D., M.P.H.; Crowell & Moring LLP

The goal of this session was to explore the problem of privacy protection in a data-rich world and consider the tensions that exist between the societal good that could come from big data research and the real and perceived risks to individuals, as well as the public's perspectives about broad data sharing. Ms. Daniel explained that speakers will consider the general landscape at present and the privacy and ethical considerations raised by current opportunities to use "big data," such as concepts of privacy, data ownership, the types and goals of data research, and how to maintain public trust.

Unexpected Forms of Risk in Data Science/Artificial Intelligence Research

- Jacob Metcalf, Ph.D.; Data & Society Research Institute

Dr. Metcalf explained that the term "big data" is a vague term used by industry, but it is not helpful or accurate. He prefers the term "pervasive data." These data are not just bigger. They bridge multiple dimensions of a person's life. They can be collected in real time and coordinated among sensors, mix varieties of datasets (public/private, research/commercial, identifiable/de-identified), and reveal intimate details about individual lives. Machines, he observed, don't respect boundaries between our different selves, but will take and analyze everything they can access. It is the nature of machine learning to "jump" domains and make predictions that apply what it knows from one domain to make an inference in another.

Most pervasive data science research uses methods and types of data that exempt it from the regulatory framework, including the requirement that it be reviewed by an Institutional Review Board (IRB). It typically uses pre-existing "public" datasets, including data borrowed/bought/gleaned from Internet services. These data are usually sufficiently de-identified to qualify as "exempt" under the Common Rule. The risks posed by these data are generally downstream, unlike the harms that IRBs operating under the Common Rule are used to considering. Also, pervasive data research does not require an "intervention" as defined by the Common Rule and involves the use of tools that are so prolific that literally everyone may be a research subject and anyone can be a researcher. Further complicating the issue of oversight is the fact that such researchers may not be associated with institutions that could provide mechanisms for ethical review of their proposed activities.

The heart of pervasive data research is the development of models that can predict behavior. The "creepy" factor, as many see it, is that these models can also be used to *influence* behavior. The findings do not apply only to the knowing or unknowing "subjects" whose data were used in developing the predictions, but also to others who never contributed data and can now be targeted with specific messages, such as political or sales pitches.

Dr. Metcalf reviewed the genesis and evolution of the "research ethics scandal" involving Cambridge Analytica (CA). Briefly, the research made use of viral quizzes on Facebook – presented as fun personality tests – to collect the "likes" not only of participants who took the quiz, but of all their friends. This information could be used to predict voting preference, among other traits. No academic research was ever published as a result of this research. The proposal for the research was approved within a day by a university IRB, pointing to the need for better understanding of the implications of this type of research among IRB members. The models developed through this research – highly portable and economically valuable – were used by Cambridge Analytica in the 2016 campaign and are still held today by CA's successor company, Emerdata.

The Cambridge Analytica episode has many hallmarks of a research ethics scandal in the age of pervasive data, which Dr. Metcalf enumerated as follows:

- Metrics jumping between domains, e.g., psychiatry to social media profiles to electoral data,
- Research that is exempt under Common Rule for narrow technical reasons,
- Blurred lines between academic and commercial research,
- Use of Application Program Interface (API) tools intended for commercial and advertising purposes to gather data for academic research,
- Abuse of mTurk workers (workers accessed through an Amazon crowdsourcing mechanism),
- Deceptive/opaque recruiting tactics for human subjects – a strong signal of unethical research,

- Predictive population models as research output become tools for intervention in individual lives, and
- Downstream effects nearly impossible to imagine because the models are highly portable and far more valuable than the actual data.

Striking a Balance: Benefit-Risk Analysis for Big Data Research

- Brenda Leong, CIPP/US; Future of Privacy Forum

Ms. Leong's remarks explored the tension between the need for beneficial research using big data and the responsibility of investigators to respect privacy and maintain public trust. There is a growing awareness of the "creepy factor" alluded to by Dr. Metcalf, and new ways of restricting such research are being explored. For example, the California legislature made several amendments to the California Consumer Privacy Act (CCPA) in September 2019, including expanding the definition of "data broker." A data broker, "a business that knowingly collects and sells to third parties the personal information of a consumer with whom the business does not have a direct relationship," must register annually with the state Attorney General, and increasing monitoring and controls by the state are possible.

In this new context, what is research? In contrast to traditional research, big data research often involves the use of data that may be collected as we live our lives rather than through a defined intervention. It is more difficult to draw a line between research and product development (for example, tools to influence behavior based on profiles). When do various legislative requirements come into play? Does this depend on the source of data, where the research is conducted, or other factors?

Ms. Leong enumerated the wide variety of harms that could result from this type of research. Individual harms (as opposed to collective or societal harms) may include loss of opportunities for employment, social benefits, insurance, housing, or education. For example, algorithms may be used to show females or other specific groups different types of job opportunities. Economic opportunities may vary, with algorithms employed in the service of credit discrimination and differential pricing. Social detriments might include dignitary harms, even surveillance and loss of liberty.

Harms may also result from unintended leakage of information. A model's output may be used to recreate the model and discern the identity of specific individuals. Behavioral harms or attacks could involve manipulating the model directly to discriminate against groups based on specific factors. Collective harms may create risks and impacts for people whose data were never included in developing the model; nevertheless, information about them can be inferred. Increasingly powerful predictions can affect those outside the model's original "ecosystem" (e.g., Facebook). For example, an application might be developed to evaluate any person's emotional state in a non-research setting without their knowing participation. "Consent" is no longer a sufficient control. A new framework is needed to assess risks. (See, for example, a 2018 white paper developed by the Future of Privacy Forum: Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models.)

Analysis of specific projects involves examining complex tradeoffs of potential harms and benefits for individuals, groups, and society. It is necessary to document initial objectives and underlying assumptions and identify undesired outcomes. Models must be evaluated to determine their purpose and assess the accuracy, transparency, fairness, and potential applications of algorithms. Review must also attempt to foresee the significance or personal impact of deploying the model on the organizations that use it, end users, third parties, individuals, and social systems. Another important consideration is whether the data produced are really accurate and reliable, since use of the models in some contexts may have serious consequences (for example, facial recognition models used to provide evidence of criminal activity).

Tools do exist for facilitating this type of analysis and providing protection to unwitting "subjects." Privacy Enhancing Technologies (PETS) can, for example, help people protect personally identifiable information (PII) online. "White Hat/Red Team" hacking exercises may be useful in identifying data vulnerability. Business processes can develop multiple lines of defense within organizations. Law and policy can also provide some controls on egregious misuse of models, as well as guidelines for what is acceptable. For example, the U.S. Department of Homeland Security has developed Fair Information Practice Principles (FIPP) to guide its privacy program. Ms. Leong assured listeners that people are working hard to find ways of monitoring trends to identify concerns and to find the best way to allow legitimate profits and benefits from this type of research without enabling exploitation.

Understanding Individual Privacy

- Cinnamon Bloss, Ph.D.; University of California, San Diego

Dr. Bloss's presentation focused on how people think about privacy. She noted that the concept of privacy rights has been around a long time and was embodied in laws in the late 1800s, partly as a result of concerns about the work of photographers and newspapers. Although we often discuss privacy as if we all understand what it means, she maintained that it really means different things to different people. Individuals differ in what information they are willing to share, with whom, and for what reasons.

To get a better understanding of how people think about privacy, the National Human Genome Research Institute funded a study with the following specific aims:

- To refine a conceptual model of privacy through literature review, individual interviews, focus groups, consultation with experts, and analyses of preliminary data;
- To develop a psychometrically sound instrument to measure individual Privacy Affinities and Privacy Environment Responses to personal health data technologies; and
- To administer the scale in a larger population and use it to explore the relationship between privacy and other factors, including the propensity to adopt Personal Health Data (PHD) technologies, the propensity to share PHD for research, and disease type and stage.

Dr. Bloss reported preliminary findings from this study. The qualitative approach used 44 interviews and 9 focus groups to explore attitudes to privacy and sharing personal health information. Some of the values that emerged reflected cultural ways of thinking about privacy that are poor predictors of individual behavior because they are too widely accepted. These included:

- "Moral right": Individuals should have control over information about themselves because that is the correct or moral arrangement.
- "Personal responsibility": Maintaining privacy is something that people must work at.
- "Tradeoff": Privacy or personal information may be traded strategically for benefits, goods, or services.
- "Nothing to hide": Privacy is only necessary to protect information that is sensitive or stigmatizing.
- "Fatalism": Privacy is already lost or does not exist.

Individual privacy values are more useful in predicting behavior. They differentiate people and show how they think about sharing or protecting information. Dr. Bloss envisioned these on a grid in which attitudes are grouped in four categories – open, intimate, anonymous, or reserved – and paired with four different ways of thinking about sharing personal health information that might be associated with each category in different circumstances, as shown below.

		Reasons to Share – Reasons to Fear (Factor 1)			
		Motivated *(many reasons to share, few reasons to fear)*	**Indifferent** *(few reasons to share or fear)*	**Conflicted** *(many reasons to share and fear)*	**Concerned** *(few reasons to share, many to fear)*
Interpersonal – Institutional (Factor 2)	**Open** *(willing to share interpersonally or institutionally)*	I am almost always happy to share my info and I don't worry about it.	I don't mind if others have info about me, but I wouldn't go out of my way to share it.	If there is not a major downside, I will share.	I'd like to share my info, but it makes me too anxious.
	Intimate *(prefer to share interpersonally)*	I need to be able to visualize the person who will benefit before I will share my info.	If someone wants to know about me they can ask me directly.	If there is a good reason, I will tell someone my story, but I'm uncomfortable with my info being in a database.	I worry about sharing my info, but when I tell people I know, I feel more in control.
	Anonymous *(prefer to share institutionally)*	I don't like to talk about myself, but I am happy to share my info with companies or researchers.	I don't mind that companies or researchers have my info if they don't expect me to do anything.	I don't like to talk about myself, but I'll share my info with researchers or companies if there is little risk.	I will only share my info if I feel sure that the info will be kept anonymous and cannot come back to haunt me.
	Reserved *(prefer not to share at all)*	I usually don't like to share, but this is a really good cause, and/or I really want the benefit.	I just prefer to keep things to myself; there is not much that can make me want to share my info.	I don't really like talking about myself, and I'm generally not going to share my info.	I don't tell anyone anything because it's not their business; it is dangerous when people have too much info about me.

Currently, the project team is in the process of developing a psychometric assessment tool with four subscales based on this analysis. The hypothesis is that subscale profiles will result in combinations that may be described as Privacy Types (similar to personality types) that differ in their openness to data sharing.

In conclusion, Dr. Bloss suggested different reasons we might seek to understand and measure individual attitudes to privacy. These included:

- To promote rigorous research on an ill-defined topic,
- To understand people's privacy-related behaviors (for example, why people might say one thing and do another),
- To enable safe data sharing for biomedical research (for example, by tailoring informed consent or developing decision aids),
- To enhance patients' control of personal health data,
- To develop approaches for addressing privacy concerns in clinical settings (for example, by tailoring telehealth interventions), and
- To promote user-centered design of health technologies and information technology.

Panel Discussion for Session I

What does giving permission mean? Ms. Daniel observed that personal preferences and profiles stored on online devices provide a framework for using and protecting data, but people do not understand the downstream risks associated with their data. Patients may not understand what giving permission to use their data actually means in this context. How do we address this problem?

Dr. Bloss called this a "huge challenge" that speaks to the limitations of the concept of informed consent in this arena. There is no easy answer. One approach is the use of decision aids to help people who may not understand intricate issues related to data management but can certainly link certain values to decisions and outcomes. Trusted entities could develop such tools.

Dr. Metcalf called, instead, for collective protections. He held that researchers should not rely on individual decision making to protect us as a community. While considering potential harms to individuals is "incredibly important," there are certain choices individuals can't make. The public health analogy is helpful here: for example, children should be vaccinated to have access to public schools. The potential for collective harms calls for collective solutions.

The challenge of unregulated research. In devising strategies to prevent harms from big data research and protect privacy, Mr. Barnes highlighted a "jurisdictional question" related to the First Amendment of the Constitution. Some research can be regulated because the Federal Drug Administration (FDA) can address commerce that crosses state lines or because the research is funded by the federal government. Even when the research is subject to the Common Rule, in some circumstances investigators can legally do what they want with their data. How do we square open and transparent use of public data with the objective of preventing harms?

In response, Dr. Metcalf observed that the fact that the Common Rule mandates the use of IRBs when federal funds are used does not mean that IRBs cannot be used in other contexts. Institutions can control the use of their own resources. They can certainly tell researchers that the researchers cannot use the institution's labs unless their research is reviewed. Beneficence and justice still matter.

When consumer data are at issue, Ms. Leong said that consumer protection laws allow companies to use data for secondary purposes with permission, but consumers are given some protection from exploitation through the agreements they make with business entities.

Does HIPAA help restrict the use of big data? Ms. Daniel observed that the Health Insurance Portability and Accountability Act (HIPAA) offers rules about data use that take collective harms and benefits into account. Certain uses of data are allowed, and the individual doesn't get to say yes or no. When this law was passed, she said, we made a set of choices as a country about what uses of data are permissible. However, the law is now 20 years old, and the issues we face regarding big data were not apparent when it was passed.

The HIPAA law, Mr. Barnes observed, is based on commerce law and governs the electronic submission of data. However, business also has rights, and the jurisdiction of the Food and Drug Administration (FDA) to suppress speech by companies about their products is currently under attack. If contracts are established and violated by a company, that is fraud – but if data are truly in the public realm, it isn't clear how we get the right to control how these data are used in all contexts. We are just not set up as a society to do this.

How do we bring local context into the consideration of privacy issues? Dr. Buchanan wondered how local context can be addressed as privacy issues are considered. In general, local context is addressed through IRB review, which takes into account the norms and values of different subpopulations. Groups may view privacy differently. How do we ensure IRBs are prepared to address varying expectations about privacy?

Dr. Bloss suggested that to address cultural issues related to privacy, as well as those that arise in other areas, diversity among investigators and IRB members is part of the solution. We can amend the ways in which we approach the issue through empirical research. However, more work is needed in this area.

How should we think about "privacy"? While we use the term "privacy" a lot, Ms. Daniel observed, we have heard that it means different things. How should we think about it? Are there different constructs to address it? What's the right framework for thinking about risks, benefits, and trust? Do individual differences make it difficult to address this?

Dr. Zimmer cautioned against "falling into the consent trap." Once collected, data are removed from context and the value of informed consent is limited. Instead, he suggested, we need to give more attention to the life cycle of the data and what happens to them over time. Since intentions change, Dr. Kilpatrick suggested, a sequential consent process is more appropriate in this type of research. She found Dr. Bloss's work in this area "compelling."

Ms. Leong held, however, that "we are way beyond the point that any individual can control the chain of consequences" once data are made available. Millions of people may be involved, and consent is no longer the main defense against misuse. Socially defined purposes are more important at this scale.

Dr. Metcalf agreed with Ms. Leong. Once artificial intelligence "goes out and does something in the world," controlling "who knows what" is very difficult. The best approach is to focus not on individual control at the beginning of the chain, but on social control at the output stage. Spelling out "who can do what to us" through a clear social contract has more potential. Mr. Gupta agreed that policies related to the use of data are more important than a focus on consent, especially given the fact that people tend not to understand what they are consenting to when they give out data, or even the fact that they *are* giving data (for example, every time they use a cell phone).

What is the right approach to "consent" in this context? Given the complexities just explored around the traditional consent process as applied to research using big data, Ms. Daniel asked, what is the right framework for thinking about the possible benefits and harms of such research?

One of the challenges in addressing this issue, Mr. Barnes noted, is that no data are ever really deidentified. Do we dispense with the notion of exemption on this basis? It is often said, and has been said here, that we should not ban the research, but we should rather try to ban bad uses of the data. These are much tangible. Mr. Gupta agreed, noting that "the data are not necessarily bad" in themselves – it is the ways they are used that may cause harm. However, Dr. Zimmer observed that we cannot confidently predict harms where pervasive data are involved.

Because of her work with rare disease foundations, Dr. Li saw this problem "through a different lens." Many people with such diseases feel strongly that data should be shared. She doubted that it was possible to clearly differentiate among types of data to understand which should be protected.

Dr. Garfinkel asked Mr. Barnes if banning misuse of data would violate the First Amendment. He responded that misuse could be prohibited, but you probably could not ban the reidentification of public data.

How do we think about harms? How do we even know what they are? Ms. Daniel said it was apparent that setting rules about how data are used seems difficult unless we are able to determine up front what the potential harms might be and how to prevent them. Also, we've heard that people think about harms differently. How do we address this challenge?

Dr. Garfinkel observed that no one has yet mentioned the usefulness of integrity models, which offer well-developed frameworks for analyzing sender and recipient interactions. One challenge in applying the models, however, is the lack of transparency and visibility around how data are collected. We need to begin by making passive collection of data visible so we have a better idea what corporations are taking and how it might be used.

Dr. Metcalf agreed that an integrity framework is useful for looking at informational harms and can be implemented so as to make it hard to make a bad decision. Platforms typically "don't report back"; if data were collected years ago and were unlabeled, the platform no longer knows to whom they once belonged. However, we are starting to see platforms that have the capacity to present metadata to scientists to enable good intent. It would help to think not just about banning inappropriate uses of data, but also about facilitating the best intent. For example, if protected health information is used, all features should travel with the metadata.

"What about reidentification?" the moderator asked. Dr. Garfinkel said deidentification does not work; anonymization does not produce anonymized data.

Mr. Barnes wondered what the real concern was regarding algorithms: that they are used to make decisions at all, the possibility that they are wrong – or that they actually work.

How do we provide training and promote cultural change among researchers? Dr. Bloss stressed the importance of training data scientists to "first do no harm." We currently have few resources to help train researchers about the specific ethical considerations for research with big data, Ms. Kasimatis Singleton said, and we need to think through what would be useful. Mr. Barnes observed that cultural change among such researchers, if it can be achieved, would be much more long-lasting than trying to find a legal solution.

Could we be overestimating the potential harms? An audience member wondered whether we might be overestimating the potential harm from big data and associated privacy violations. Perhaps there is a cultural shift in the way people think about privacy that we should take in account.

When people release data on the Internet about themselves, Dr. Garfinkel observed, people are usually taking advantage of a platform and assuming it has certain controls. They may not realize their data are being archived, and they typically can neither see nor control the flow of information. There is a role for education and a role for regulations, but there is also a system development issue. Mr. Gupta noted that some people share too much, and others do not even know they are sharing. Ms. Leong added that people who never supplied their data are also potentially exploited and harmed. Dr. Li reminded panelists that Dr. Metcalf's presentation showed how our own choices may have implications for potential harms to our friends. Science tends to race ahead of regulatory frameworks, and that is happening now.

Dr. Garfinkel wondered how the process we are discussing differs from public education: we collect data, build models, and manipulate them as a society. Dr. Metcalf rejoined that public education is subject to democratic control, and people who do not like the choices made in public schools often have other options, such as home schooling or private schools.

Ms. Leong observed that the traditional model of education was the teacher in the classroom, collecting data on each individual student and knowing each person. Now we have companies coming into the education field collecting information on thousands of students and handing tools to teachers based on their analysis. Some people think this is dangerous, while others are enthusiastic and believe it can be helpful.

Ms. Daniel closed by observing that the same applies in the world of health care. The doctor now has access to tools and information far beyond the scope of the doctor's direct knowledge of the individual in the office – but if the models really do not apply to that individual, they can cause harm.

Session II: Approaches to Protecting Privacy & Confidentiality

- *Moderator:* Mark Barnes, J.D., LL.M.; Ropes & Gray, LLP

Session II Introduction

- Mark Barnes, J.D., LL.M.; Ropes & Gray, LLP

Mr. Barnes explained that the session is intended to explore some practical approaches to protecting privacy and confidentiality, especially in health research. It addresses the challenges of privacy protection for health-related big data research conducted on a variety of platforms and in various settings. Speakers representing diverse stakeholders in the research enterprise discuss policies, techniques, and technologies for controlled use, data protection, and informed consent as ways to protect individual privacy and data confidentiality.

Transforming Health Measurement and Care Delivery Through Patient-Generated, Permissioned Data from Daily Life

- Deborah Kilpatrick, Ph.D.; Evidation Health

Dr. Kilpatrick is the Chief Executive Officer (CEO) of Evidation Health, which connects directly with millions of individuals in the U.S. and provides those individuals with the opportunity to participate in research, under informed consent, utilizing connected device data streams that they have expressly chosen to share with Evidation. Evidation works with medical product companies that sponsor studies on the Evidation platform; these studies use streaming data from connected technologies and wearables and involve both consumer and clinical grade APIs, for the purpose of measuring outcomes in the real world outside the clinic. She focused her remarks on how one private company tackles privacy and confidentiality concerns with data collection, storage, and controlled access of data generated through mobile technologies for research use.

Historically, health outcomes have been assessed using limited data from within the clinic, as opposed to using data from daily life. These data could not be accessed in the past, but now, in the digital era of medicine, it is possible to view and measure health and disease through the lens of daily life. Individuals are able to track their data from daily life and give permission to use that data in research. These data may

include their interactions with technology, their movements, their stress levels, their sleep patterns, their diet, and more. For example, a "behaviorgram" can be generated for an individual– a "snapshot" of how these connected device data sources directly reflect and measure individuals' health-related activities and behaviors in a 24-hour period. Recent study has shown that such passive, continuous measures can help distinguish those diagnosed with Alzheimer's Disease from those with mild cognitive impairment (Chan et al. 2019). Similar studies using connected device data streams and "digital assays" may be able to detect early cognitive decline, provide information on daily movement patterns to monitor surgical recovery, or quantify daily patient symptom patterns important to battling the opioid crisis. Work is underway in all these areas.

It is possible to construct rigorous clinical trials that use digital data to look at issues differently from traditional trials. Studies can involve more diverse participants with less burden on patients, who no longer need to come to a specific building to be studied. Today, more than a million people in the U.S. are participating in registries, trials, and health-related research on the Evidation platform. This makes it possible to do a randomized study with hundreds of thousands of participants.

As an increasing number of entities seek to use digital data to profile individual health and construct new biometric identifiers, Dr. Kilpatrick stressed, it is critical that we place individuals in control of how their data can be used. Individual privacy rights should be fortified with appropriate legal and regulatory frameworks that can prevent and manage attempts to discriminate inappropriately based on this information. Attention must be given not just to how data are collected and shared at the point of collection, but how they are actually used over time.

Evidation accepts the definition of information privacy used by the International Association of Privacy Professionals: the right to have some control over how your personal information is collected and used. Evidation honors this right by only collecting an individual's data with their explicit permission, seeking to use the minimum amount of data required to fulfill a specific purpose, and never providing or sharing individuals' data with clients without their consent. Its protocols undergo review by external IRBs, using cross-functional internal review to further address issues around ethical use of data. The consent process is specific and unambiguous: individuals must make an active choice to participate. Evidation also uses sequential consents that allow individuals to determine whether their data can be used or reused on specific client projects over time. Any individual can withdraw consent at any time for any reason. This means that the only data Evidation has on its platform about individual members were directly permissioned by those individuals for collection or directly provided to the company by them.

The company also takes its responsibility for secure storage of personal data very seriously. When data are stored or shared, these data are encrypted both at rest and in transmission. Access is provided on a "minimum necessary" basis and carefully monitored with internal access controls that are routinely audited. Evidation also advocates publicly for appropriate legal and regulatory frameworks to protect individuals against potential abuses of new forms of health data.

In 2012, the first individual in the US connected to the Evidation platform and provided permission for the company to collect digital data from their "wearables." Today, these data are enabling novel forms of health measurements to be constructed passively, continuously, and with individual permission. Many therapeutic areas stand to benefit.

CMS Data Products

- Andrew Shatto; Deputy Director, CMS Office of Enterprise Data & Analytics, Centers for Medicare & Medicaid Services

Mr. Shatto began by highlighting some of the many laws and regulations that govern any release of data by the Centers for Medicare and Medicaid Services (CMS). CMS is a covered entity under HIPAA, which imposes restrictions on the release of protected health information (PHI). The Privacy Act of 1974 also applies. Both HIPAA and the Privacy Act, which governs the use of personally identifiable information (PII), classify specific uses of data that are allowed without obtaining patient consent. Permitted uses of data under the Privacy Rule include for treatment, payment, and healthcare operations, as well as activities with public interest and benefit (such as health oversight). However, disclosures must be limited to the minimum data necessary. Both HIPAA and the Privacy Act also require CMS to track data disclosures, which it does through the mechanism of a Data Use Agreement.

Any public-facing data product (such as a dashboard, public use file, or interactive tables) also must be compliant with HIPAA's de-identification rules and the CMS Privacy Policy. HIPAA establishes two methods of de-identification:

- *Safe harbor* involves the removal of specific variables; or
- *Expert determination* involves a statistician certifying that the data set could not be used to re-identify an individual.

HHS's Office of Civil Rights has published guidance on de-identification, including details about how to apply both methods.

CMS makes two types of research data files available to researchers. These include:

- Limited Data Set (LDS) files, which exclude specific direct identifiers, including name, address, Health Insurance Claim (HIC) number, social security number, date of birth, and ZIP Code; and
- Research Identifiable Files (RIFs), which are custom CMS data extracts that may contain direct beneficiary identifiers.

LDS files are easier to request (they require less documentation and CMS review) and typically cost less than RIFs; however, users face additional limitations on the use of the data.

CMS has established mechanisms for supporting cutting-edge health care research:

- The Research Data Assistance Center (ResDAC) provides assistance to researchers interested in using Medicare and/or Medicaid data; and
- The Chronic Condition Warehouse (CCW) is designed to support external researchers as well as internal CMS research and analytic functions. A unique beneficiary ID allows data linkages across all CMS data housed within the CCW.

Data can be disseminated for research purposes in two ways:

> *Option 1: Providing Physical Data.* This requires files to be created, encrypted, and copied to portable media by CMS, then shipped to researchers, who must ensure the security of the data at the researcher's site. The researcher is responsible for the security and appropriate use of the data. This approach may be beneficial for users that are requesting a limited amount of data files for a

small cohort, for researchers that are working with non-CMS data sources that cannot be uploaded to the Virtual Research Data Center due to licensing or other restrictions, and for researchers that are using specialized software and/or tools. However, it is time consuming and costly.

Option 2: Virtual Research Data Center (VRDC). CMS developed the VRDC to meet researchers' evolving needs. It offers a secure and efficient means for researchers to access, analyze, and manipulate the vast store of CMS data virtually from their independent workstations. Researchers can only download aggregate results from their analyses.

Please see the Resources section at the end of this document for internet links to key CMS sites mentioned in this presentation.

The Vivli Experience in Sharing Clinical Trial Data Globally

- Rebecca Li, Ph.D., Vivli

Dr. Li represented Vivli, a nonprofit organization that offers an independent global data-sharing and analytics platform for clinical research data. Dr. Li explored the challenges encountered and solutions implemented for protecting privacy and confidentiality of clinical research data on a shared platform that operates across geographical boundaries.

The speaker described the platform as "user-friendly, secure, and state-of-the art." It uses modern tools and technologies, serves the international community, and is capable of including trials from any disease, country, sponsor, funder, or investigator. The organization provides expertise to the biomedical industry, nonprofit funders and foundations, government funders and regulators, and patients or patient advocates to help them move toward a culture of data sharing with robust security mechanisms. With 4300 trials on the site to date, representing more than 2 million participants, Vivli is currently the largest global data sharing platform.

Dr. Li reviewed the history of requirements and policies for transparency in clinical trial data (see her presentation). While sharing summary data is not as sensitive as patient-level data, sharing patient-level data requires controls. Vivli provides a secure enclave through which this level of sharing can be done responsibly.

The speaker observed that the workshop has so far stressed the importance of protecting participant privacy and given less attention to the benefits of data sharing. She argued that such sharing maximizes the value of the trial data collected and thereby respects participants' contributions to research. Depending on the type of data to be shared and the level of protections required, platforms may offer open access, managed access, or restricted access. Restricted access means that only those that provide data are able to use it or invite others to do so. Vivli respects the review process of each data contributor and has built in flexibility to accommodate various review processes into the current system. In areas where harmonization is critical for the user experience, it provides specific administrative mechanisms to accomplish this. For example, researchers are bound by a Harmonized Data Use Agreement (DUA), through which they agree to adhere to a research plan, to make reasonable efforts to publish, and not to re-identify participants.

Vivli's data request and access process allows its users to:

- Search the Vivli platform for information about available studies.
- Request an Individual Patient Data (IPD) data package. Each Data Request is reviewed according to the contributor's publicly stated requirements.
- If the request is approved, the user can access data in Vivli's secure research environment, or download it with permission.
- Analyze the data using robust analytical tools provided on the site; the user can combine and analyze multiple data sets.
- The user may disseminate research findings, which are assigned a Digital Object Identifier (DOI). The platform may be used to meet publication requirements.

Vivli's platform rests on "pillars of security" that include the signed data use agreement, the security provided by platform itself (for example, data anonymization and a secure environment for analysis), and the researcher's concern for his or her own reputation.

The Use of "Differential Privacy" as a Statistical Method for Protecting Confidentiality in Data Publications

- Simson L. Garfinkel, Ph.D.; Senior Scientist, Confidentiality and Data Access, US Census Bureau

Dr. Garfinkel explained that the mission of the US Census Bureau to collect information from the American people and to publish statistical information without compromising the confidentiality of its data subjects. It is legally prohibited (by 13 U.S. Code § 9) from making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

It is easier to reidentify personal data than you might imagine, the speaker stressed. One way to avoid this possibility is by suppressing some of the data. However, if there is a small number of contributors (say people in a particular city block) and you suppress one person's data, it is still possible to back out the data using basic algebra and identify individuals. Even if an entire row of data is suppressed, external data (such as newspaper articles) may help someone reidentify individuals. This is called "data-based reconstruction." If enough summary statistics are published, he cautioned, you can always get the original data back. This means something more mathematically powerful is needed to protect respondent data.

The requirement to maintain confidentiality has forced the US Census Bureau to adopt Differential Privacy (DP) as its new technique for statistical disclosure limitation. Differential Privacy is best understood as a system for adding carefully controlled noise to statistical products to create uncertainty, making it impossible to back out original data. It is called "differential privacy" because it mathematically models the privacy "differential" that each person experiences from having their data included in the Census Bureau's data products, compared to having their record deleted or replaced with an arbitrary record. Differential privacy offers:

- Mathematical definition for privacy loss caused by a data release,
- No dependence on attacker's external information,
- No dependence on attacker's computer power, and
- Composable privacy-protection mechanisms.

The use of differential privacy sets provable bounds on the maximum increase in privacy loss caused by the data release. However, it comes at a cost.

Differential privacy uses algorithms that allow policy makers to manage the trade-off between accuracy and privacy. The more noise, the greater the privacy; but also, the more noise, the less accuracy. Differential Privacy forces data providers to confront the reality of the accuracy/privacy-loss trade-off. It is like a knob that can be set to decide how much privacy loss is allowable. Once done, it is mathematically impossible to undo the addition of the noise to reliably get the original data. Someone could still try to use algebra to provide a solution – but it probably won't be the correct one.

There are two different approaches for using DP. These include:

1. An Interactive Query System. This is the easiest way to apply DP, but every query must be tracked and we don't know how to make all queries private efficiently.

2. Use DP to create a model that produces synthetic microdata. This approach allows microdata to be published and used indefinitely without additional privacy loss. However, sophisticated modeling may be needed to create accurate microdata.

   A special case of synthetic microdata is running DP in the "local model." DP is applied directly to the microdata. It's called "local" because people can apply the noise themselves, before sending the data to a trusted curator. This approach is easy to do, but it produces fairly poor data for a given level of privacy loss.

Additional challenges and limitations are associated with DP. Today, it works poorly with linked data tables with one-to-many relationship, geospatial information, and time series. It also does not work with imagery (such as photos) or genetic information.

Panel Discussion for Session II

Mr. Barnes observed that big data has been helpful in many ways to epidemiology and advances in health, and some even like the targeted ads. Many of those who use the CMS data base for research take the point of view that, given the tremendous benefits, people with concerns about privacy should "get over it." Is there a problem?

Should "microdata" be released? Mr. Barnes asked Dr. Garfinkel how he would handle a request from an agency to release microdata. He said his preferred approach would be either not to release it or to allow it to be released in a controlled environment. The second strategy would be to use the data to build a statistical model and then use the model to make synthetic microdata.

Should genomic data be considered identifiable? A member of the listening audience asked whether genomic data should be considered identifiable or nonidentifiable. Ms. Kasimatis Singleton suggested that in light of revisions to the Common Rule and the new definition of identifiable information, this needs to be considered carefully. There are various ways to argue that select genomic data may not be identifiable, but more analysis is needed to reach a conclusion. She wondered if it would be possible to develop a tool that could reach the conclusion, "based on what you are doing, this specific genomic data is (or is not) identifiable."

What ethical considerations apply to data sharing? Mr. Barnes invited comments on the ethics, practicalities, and legal considerations in having one institution serve as custodian for "big data" sets. If one institution has a completely identifiable data set and many researchers want to "crunch the data," what kind of safeguards are needed? Ms. Kasimatis Singleton said that gatekeepers are needed in the

middle to verify that parties receive only a limited data set. Also, the recipient of the data must agree to use it only within clearly stated limits.

Dr. Kilpatrick explained that if partners wish to use a CMS data set, they first must convince CMS of the potential benefit to participants. Then the IRB must approve the protocol. Each of the companies involved will have systemic controls in addition to the regulatory controls of pharma partners. There are guardrails that apply.

The National Institutes of Health (NIH) has a number of data bases or depositories, some of which, Dr. Garfinkel said, link to data held by the U.S. Census Bureau. (Clinicaltrials.gov holds only summary level data.) Mr. Barnes observed that NIH requires that even when deidentified data are submitted to NIH, subjects must be informed that their data will be contributed to NIH. This is also true of genomic research. NIH has gone so far as to charter a new IRB to have jurisdiction over deidentified data and is developing operative principles to address the potential for stigmatizing results.

A member of the listening audience queried, "If our institution lets another use its data and then finds they have violated the terms of use and passed it on to someone else, what should happen? How can this be prevented?" The listener asked Mr. Shatto what CMS does to prevent violations of the terms of use and whether it has experienced broken agreements. Mr. Shatto said there is a "huge risk" for academic institutions that violate terms of use. When CMS discovered that one university professor using the data also had a private business that could benefit from the data, that professor ultimately lost tenure. "We have controls."

What controls are appropriate for targeted health interventions? Dr. Metcalf noted that health care interventions targeted to individuals by using pervasive data have more protection than other categories. CMS has some controls in place, and FDA regulations may apply in some instances.

Dr. Garfinkel wondered whether it was possible, for example, to look at filtered information from Instagram and identify individuals who are depressed. If so, how could that information be protected? Could social media make it possible to target gun sales to suicide risks? Dr. Metcalf was not sure. He commented that many sources of data lack protections of any kind. Commercial providers could argue that they are not doing research and not offering medicine. Instagram could choose to block certain kinds of advertising based on filters.

Artificial intelligence (AI) can discover patterns in social media posts a human analyst could never find. Humans using social media apps may share their status as depressed persons without realizing it. Dr. Metcalf said he would prefer that these types of targeted health interventions did not exist at all because of the challenge of providing appropriate controls.

Mr. Gupta observed, on the other hand, that Facebook uses algorithms to detect posts from people who may be suicidal, and Apple's "Siri" is able to identify and respond to some suicidal statements. The capacity for large-scale analysis of posts with appropriate interventions could also be used to address drug abuse or help people overcome their diseases. Fines might be levied to prevent abuse of information, but we should not lose sight of the potential for good.

Mr. Barnes asked whether Dr. Metcalf would prohibit AI from learning something that is generalizable but would not target individuals for interventions. Dr. Metcalf observed, in response, that the "line between intervention and research is incredibly thin." The model that results from research is the exact same model you would use for the intervention. Platforms can be built to make this transition happen with ease that seems almost magical.

Mr. Barnes observed that there are many examples of research based on AI that yield health-related data. One example is the capacity to look at retinal scans to identify diabetic retinas. Dr. Metcalf said the key question is how this knowledge is used. A camera in the mall might identify someone as a diabetic in an effort to find people who love sugar. It is not the case that corporations cannot access tools that benefit the medical profession for corporate gain. In short: "The implicit distinctions we relied on to be proxies for our values don't work anymore."

How are academic and commercial health research different? Mr. Barnes observed that there is a tendency to assume that academic health research is good and commercial research is bad – but how are they really different? Academic institutions may have commercial interests. Many research institutions are allowed under HIPAA to give out limited data sets for research purposes and no distinction is made between academic or commercial entities.

Ms. Daniel said the question of whether limited data sets can be given to commercial entities comes up frequently. With an appropriate data research agreement, it should be possible to ensure there is generalizable research that has the capacity to advance the field, but the distinction between commercial and academic interests is definitely "muddy."

Many institutions act as gatekeepers for data that can be valuable to many potential users, Ms. Kasimatis Singleton observed. They need to consider the best uses of these data – not simply what you are allowed to do, but also what is ethical to do.

Mr. Barnes said this type of analysis hinges on cultural issues. Transparency can be seen as an ethical requirement. For example, the European Medicines Agency (EMI) requires companies that apply to market their products in Europe to make patient-level data available to anyone, either through their own company or through a data sharing platform. The EMA requires complete transparency even in small trials, such as those involving rare diseases. Experience suggests that these data are used primarily by other industries that want to conduct their own trials and benefit from competitors' research. Different worlds protect privacy and transparency.

Dr. Kilpatrick observed that we simply talk past each other as we explore the tension between openness and privacy. It is time to consider what kind of framework can be put in place with appropriate balances to protect both.

How should researchers communicate with the public about how data are used? A member of the public asked what kind of information should be given to people who have consented to downstream uses of their data. Dr. Garfinkel observed that to have meaningful informed consent, people need to know who is using their data and be able to withdraw it. This is completely achievable with today's technology, but companies don't want to do it. Dr. Kilpatrick noted that informed consent documents are already 25 pages long on the average, so she would be concerned about adding additional information about data sharing.

Mr. Shatto observed that many people do not read Terms of Service (ToS) for apps, such as those that collect health information. It would be helpful to have those written in plain language so people know what they are agreeing to. Dr. Garfinkel said that many studies have suggested that this information could be presented more effectively – for example, in a manner similar to the way nutrition labels work. Currently, the education level required to understand ToS is high. Ms. Daniel and Mr. Gupta agreed, arguing that a standardized label could readily show what is and is not covered by the agreement. Mr. Barnes wondered if the concept of "key information" in the new Common Rule could be extended to privacy policies.

Dr. Bloss saw a more hopeful approach in educating children at an early age about appropriate online behavior and helping them think about the kind of information they want to put out on social media.

Ms. Kasimatis Singleton observed that even if we find successful ways to communicate with subjects about what will be done with their data, we are still falling short in explaining *why* someone would want to use their data. The field has an obligation to educate the public about how their data may be useful in research and the importance of validating results.

Mr. Barnes observed that there is debate about whether there should be adaptive approaches to informed consent so that people are reconsented periodically and whether the process should be the same for deidentified and identified data.

What ethical considerations apply to commercial use of health data? A member of the public observed that in a clinical setting, patients are asked to share their data with their insurance company so the insurance provider can refine its products. How is that different from researchers' use of big data, or from Instagram or Facebook identifying people with depression? Isn't all of that human subjects research? Dr. Zimmer said he has often heard the question from researchers: "If Facebook does this, why can't I?"

Mr. Barnes suggested that while "big data" research has relied traditionally on CMS claims data, Census data, and other public data, these large data sets are now dwarfed by data collected commercially. How, he asked, can we do relevant research without access to commercial data bases? How can academic researchers get access to commercial data bases to derive academic knowledge? Of course, industry is not required to give anyone access to their data, and if they do make an agreement they want to see any potential publication before it is submitted and ensure they maintain ownership of any algorithm the researcher might develop. Dr. Garfinkel added that corporations are also afraid that researchers might leak information that embarrasses the corporation.

Session III: Protecting Privacy & Confidentiality: A Shared Responsibility

- *Moderator:* Elizabeth Buchanan, Ph.D.; University of Wisconsin, Stout

Session III Introduction

The goals of this session are to discuss some of the challenges facing IRBs and institutions in the review of big data research that falls under the Common Rule as well as in ethical oversight of big data research that falls outside the scope of the Common Rule. Invitees explore the possible roles a variety of stakeholders may play in supporting responsible conduct of research involving big data.

IRBs and Big Data Research

- Michael Zimmer, Ph.D.; Marquette University

Dr. Zimmer described empirical research with IRBs that provided insights on how IRBs review big data research.

The speaker observed that the growing use of pervasive data in research is creating new conceptual gaps in our ethical understanding of:

- Privacy and anonymity
- Informed consent, and
- Harm and human subjects.

Challenges are complicated by the increasing ease of doing big data research. Also, computational social science includes disciplines and practices that are not always focused on "human subjects" as we have understood them. Tools to engage in this kind of work are generally available and new groups of investigators are working with data. As a result, there is often a lack of ethical training or proper oversight. This type of research is not always conducted within an "IRB culture." The speaker urged a closer connection in institutions between computer sciences departments and IRBs.

A 2009 study sought to provide empirical measures of ethical threats, concerns, and approaches across stakeholder communities (Buchanan & Ess, 2009). The study surveyed 334 US IRB members and sought to answer the following questions:

- Do IRBs feel suitably prepared for projects that rely on "pervasive data"?
- How are IRBs reviewing "pervasive data" projects?
- What are the key issues that determine whether "pervasive data" projects undergo full review?

At that time, investigators found:

- Less than half indicated internet-based research was "an area of concern or importance" at that time;
- Only 6% of respondents had guidelines or checklists in place for reviewing internet-based research protocols; and
- 18% said guidelines or checklists were under development.

A more recent study surveyed 59 US IRBs in 2015 (Vitak et al., 2017). While the majority of IRBs in this study saw ethical issues associated with the research, results were still concerning:

- 93% of respondents agreed that there are ethical issues unique to research using "online data."
- However, only 55% said they felt their IRBs were well versed in the technical aspects of online data collection, and
- Only 57% agreed that their IRB had the expertise to stay abreast of changes in online technology.

Pervasive Data Ethics for Computational Research (PERVADE) conducts research in this area with support from the National Science Foundation (Grant #1144934). It is currently surveying IRB members recruited via the PRIM&R blog, IRBForum, direct email, and social media to better understand their work with pervasive data and their approach to project review. The project defines "Pervasive Data" as "rich personal information generated through digital interaction and available for computational analysis." It refers to research that:

- Gathers digital data about people;
- Uses computational methods to understand individuals' or groups' health, habits, routines, or beliefs; and
- May frequently (but not always) collect data without the awareness of the studied population.

To date, the survey has received 79 valid responses, of which 80% were from a college or university. The majority of respondents (63%) reviewed only nonmedical protocols; 30% reviewed both medical and non-medical protocols. Basic findings are reviewed below. (See presentations for more comprehensive results.)

- *How many protocols involving pervasive data does your IRB review annually?* About a third had 50 or more protocols with this type of research in a year.

- *Are IRB members at your institution well-versed in the **technical** aspects of the collection and use of pervasive data?* Only 25% feel their IRB grasps the technical aspects of pervasive data protocols, compared to 55% in 2015 who agreed when asked a similar question about online data. The speaker suggested that this is a gap that needs to be addressed through education.

- *Are IRB members at your institution well-versed in the **ethical** aspects of the collection and use of pervasive data?* While only 25% feel their IRB understands the technical details, 50% are confident their IRBs grasp the ethical issues.

- *Does your IRB have a specific checklist, review tool, policy or guidelines for reviewing protocols that rely on pervasive data?* 76% do not, but 18% have such a resource under development. Only 6% already have such guidelines. This means, said Dr. Zimmer, that we're in the same situation as IRBs were in 2007-2008: lacking specific tools to help address internet-based research protocols.

- *In reviewing protocols relying on pervasive data, which regulations or guidelines do you consult or rely on, if any?* The most frequently used sources are the Common Rule, the Belmont Report, and SACHRP recommendations.

- *Does your institution provide specific training sessions for **researchers** that address the ethics of the collection and use of pervasive data?* Fourteen percent of respondents said such training was required, and 11% said it was optional. Dr. Zimmer observed that this is better than 2007-2008, when only 5% required such training and 9% said it was optional. However, 65% of respondents still do not offer any specific training for researchers.

- *Does your institution provide specific training sessions for **IRB members** that address the ethics of the collection and use of pervasive data?* Nineteen percent of respondents required such training, and 8% said it was optional. This is about the same as 2007-2008, where 19% required such training and 6% said it was optional. The majority of IRBs (70%) offer no training specific to pervasive data.

The PERVADE team then presented respondents with various hypothetical pervasive data research protocols and asked how their IRB would likely categorize this project, what level of review would be needed, and what the key factors would be in making this determination. For many hypotheticals, there

was general agreement on how their IRB would likely review the protocol. The "publicness" of the data tended to be the primary factor, more so than whether there was consent. However, if the project involved a sensitive topic, or data were being co-mingled, then consent appeared to matter more and additional review was more likely to be triggered.

For many hypotheticals, there was considerable disagreement on how their IRB would be likely to review the protocol. Terms of service were a key indicator, and violations of the terms of service were said to trigger a full review at some, but not all, IRBs. Some felt that the institution's attorney – but not the IRB – should be informed in such a case.

In summary, Dr. Zimmer reported that pervasive data scenarios present a "mixed bag" of responses. Most approach "public data" either as not human subjects research or as exempt, with or without consent. Inferring mental health or sexuality triggers more rigorous review, and lack of consent in more sensitive protocols can trigger confusion. Key takeaways from the presentation include:

- The PERVADE Pervasive Data Ethics project continues to work to measure IRB attitudes and readiness for "pervasive data."
- PERVADE will continue to process, and publish, results of the IRB survey.
- Focus groups will be held at the PRIM&R Advancing Ethical Research (AER) conference.
- PERVADE plans to take a "deeper dive" into how some IRBs are reviewing pervasive data protocols.
- PERVADE plans to hold stakeholder workshops and will be developing a toolkit to help IRBs navigate these complicated issues.

A Framework for Ethics Committees for Reviewing Research Protocols with Privacy and Confidentiality-Related Risks in Electronic Environments

- Adarsh K Gupta, D.O., M.S., FACOFP; Rowan University School of Osteopathic Medicine

Mr. Gupta sought to present a practical and comprehensive framework that ethics committees could use when reviewing research protocols with privacy and confidentiality-related risks. He cited three reasons that it is important to secure research data:

- To comply with regulations,
- To protect the individual subject's identity, and
- To protect the integrity of the research institution and the researcher.

Securing digital information comprises the following standards:

- *Confidentiality* – Information is not disclosed to unauthorized entities;
- *Integrity* – Information is not altered or destroyed in an unauthorized manner and is transmitted accurately;
- *Availability* – Information is accessible and useable upon demand by authorized persons.

Mr. Gupta reviewed the various federal regulations pertinent to protecting the privacy of health information (see presentation) and observed that individual states also may have additional regulations on protecting human subjects information. He also highlighted less familiar regulations:

- *Export Administration Regulations (EAR):* These regulations govern export-controlled research that includes information that is regulated for reasons of national security, foreign policy, anti-terrorism or non-proliferation. These data cannot be stored on systems outside the United States.

- *Student Educational Records or FERPA Family Educational Rights and Privacy Act:* These regulations pertain to records that contain information directly related to a student (such as grades and transcripts) that are maintained by an educational agency or institution.

The speaker reviewed the various kinds of risks that may result from breaches of confidentiality:

- Risk to human subjects of identity theft, embarrassment, misuse of personal information, or victimization in fraudulent scams;
- Risk to research of loss of data and loss of integrity;
- Risk to the research institution of loss of trust, media attention to security lapses, litigation by subjects, penalties, and prosecution;
- Risk to the investigator of loss of data, time and money; embarrassment; media attention to security lapses; litigation by subjects; internal disciplinary action; penalties; and prosecution.

Mr. Gupta stressed the uniqueness and the multiple security challenges associated with mobile devices. People may not be aware that the data on such devices is being viewed and even stolen. Mobile devices can also be stolen, hacked, or infected. He provided a variety of specific suggestions for protecting data in mobile environments, including:

- Creating a strong PIN or passcode,
- Enabling the option to "erase data" after certain number of failed attempts,
- Updating apps regularly,
- Turning off services when not in use,
- Encrypting sensitive data,
- Backing up data,
- Putting an email address or an office phone number on the device so if it is found it can be identified and returned to you, and
- Installing an app that allow you to find, lock, or wipe your phone remotely if lost or stolen (for example, find my iPhone, Lookout, Lost Phone, or Autowipe.)

He further cautioned that even reputable companies can make mistakes and leave security loopholes wide open for hackers. Many apps may not be reputable and will steal data. Also, many do not use encryption properly (or at all) so passwords and other data may be exposed, especially when using wireless. It can be safer to use a browser and access a secure web site than to use a dedicated app.

Mr. Gupta also highlighted specific security concerns associated with commercial cloud storage:

- The cloud acts as a big black box; nothing inside the cloud is visible to the users, and users have no idea or control over what happens inside a cloud.
- Even if the cloud provider is honest, it can have malicious system administrators who can tamper with the data and violate confidentiality and integrity.
- Clouds are subject to traditional data confidentiality, integrity, availability, and privacy issues.

For financial reasons, many cloud providers locate some of their servers outside the United States. Because you do not know the physical location of the servers on which a provider stores your information, it may not be appropriate to use commercial cloud providers for certain research applications.

The speaker cautioned that when human subjects research involves the collection of private information or a promise of privacy or confidentiality to research participants, IRBs should not assume that Dropbox.com—or any cloud provider—is a secure environment for such data. Also, research data with restrictions on the participation of foreign nationals, restrictions on publication (prior approval or prior review), or restrictions imposed by nondisclosure agreements should not be stored on a commercial cloud service.

In summary, the speaker offered the following considerations in review of research protocols:

- Be cognizant about the type of data storage option used by the researcher – cloud, USB or other portable device – and its security risks;
- Be cognizant about the use of mobile devices by researchers and their security risks;
- If possible, offer options to provide a secure e-environment; and
- Limit storage of sensitive information on e-environments as much as possible. If you put it there, you are liable for the consequences.

Facing the Future: Operational Solutions to the Regulatory Challenges of Big Data Research

- Megan Kasimatis Singleton, J.D., M.B.E., CIP; Johns Hopkins University School of Medicine

Ms. Kasimatis Singleton noted that institutions face complex regulatory challenges in the fast-evolving field of data science. Her presentation sought to:

- Outline key operational challenges for IRBs in reviewing big data research, and
- Identify potential practical solutions IRBs/organizations might leverage to address these challenges, including institutional operational processes and policies to uphold their commitment to ethical research.

The first challenge with "big data" research is determining whether it is subject to IRB review. When research with big data is subject to IRB review, IRBs must consider difficult questions:

- *Waivers of Consent:* When is Informational Risk = Minimal Risk?
- *Minimum Necessary:* Is the data to be used the minimum necessary to accomplish the research objective? In the context of big data research, this can be especially challenging.
- *Sufficient Expertise:* Must the nature and composition of IRBs change to ensure there is sufficient expertise to evaluate the research being proposed? IRBs may not have a pool of experts on whom they can draw.
- *Fixed vs. "Dynamic" Approval:* How can IRB review processes accommodate the exploratory nature of big data research? Such research may be complex and may not fit into the typical review structure. The project must be described in such a way that the IRB can understand it. How do we get to point that we can confidently approve such research, given its complex and dynamic nature?

The speaker offered four operational solutions to address the challenges posed in reviewing this type of research and explained how Johns Hopkins University School of Medicine is implementing these solutions.

Operational Solution 1: Create the Minimal Risk Environment. This solution requires offering organizational solutions in which researchers may interact with larger scale data in a secure environment. Johns Hopkins offers a SAFE Virtual Desktop where investigators can access and analyze data within a secure environment and has created a Precision Medicine Analytics Platform (PMAP) to enable larger scale exploration of data on an institutionally managed platform. As described on the Hopkins website, PMAP "pulls data from the Epic Medical Record and other data sources into a Data Commons, where the data are integrated together and available in a format that is operable by sophisticated machine learning and natural language processing technologies." Additionally, Johns Hopkins Medicine has created a "gatekeeper" function through the use of trained data stewards who provide an added layer of security by addressing data queries and ensuring the use of data is consistent with the approved protocol.

Operational Solution 2: Steer Researchers to Solutions that Satisfy Confidentially Concerns. Johns Hopkins Medicine has taken specific steps to help steer researchers to best practice solutions for data storage and management including:

- Piloting use of a data security profile and checklist.
- Creation of a risk tier matrix for data storage and management that categorizes planned use and management of data into specific designated tiers to identify the data risk associated with the proposed research. Building off of this matrix, specific triggers for organizational data review have been identified and incorporated into the eIRB system.
- A Risk Tiers Worksheet is now embedded in the eIRB system and required for each application. The worksheet consists of different layers that capture key information such as the type of data being collected and the plan for storage. Using the matrix, it is possible to create a ranking and see what levels of review would be triggered by the proposed plan.

This system is intended to help investigators planning big data projects learn which data storage and management solutions constitute best practice and achieve a lower level of risk (for example, by choosing a more secure storage option).

Operational Solution 3: Partner with the Appropriate Expertise. At Johns Hopkins, designated IRB representatives sit on a research data sub-council of the Johns Hopkins Medicine Data Trust. When review by the Data Trust is triggered, Data Trust approval is required prior to final IRB approval. The Data Trust has been incorporated as a formal ancillary committee within the eIRB system with its own dedicated review workspace to help facilitate the review process. Representatives of the IRB, the Data Trust, and Information Technology (IT) meet regularly to improve processes and workflow.

Operational Solution 4: Accommodate the Changing Environment. Johns Hopkins realizes that researchers need to be able to explain the processes they are using more clearly, and traditional protocol templates may not accommodate project design for big data research. The institution has changed some traditional forms to ask different types of questions more appropriate for this type of research. The institution is also considering flexible options for ensuring oversight, such as more frequent progress reports, a trained data steward as a study team member, and independent assessment to determine that the data requested is needed to answer the research question.

Ethical Considerations for the Review of Big Data Research Beyond the Common Rule

- Brenda Leong, CIPP/US; Future of Privacy Forum

While IRBs are typically focused on informing subjects *before* they participate and the data are collected, new research opportunities may involve the use of data already gathered by entities not subject to the Common Rule. Organizations and researchers that fall outside traditional frameworks for review are increasingly encouraged to pursue internal or external review mechanisms. To prevent unethical data research or experimentation, experts have proposed a range of solutions, including the creation of "consumer subject review boards," formal privacy review boards (Calo, 2013), private IRBs (Polonetsky, Tene, & Jerome, 2015) and other ethical processes implemented by individual companies (Schroepfer, 2014).

In reviewing big data research, IRBs need to consider:

- *Is it research?* If so, what kind? Will it advance general knowledge? Does it involve systemic investigation, including testing and evaluation? Is it intended for publication? Is it reproducible? Will data be shared/made available? Is it Intended to improve, maintain, or add new features or create related products?

- *Is there identifiable private information?* If so, where did the observation take place? Was it based on direct or indirect interaction with subjects? Did subjects expect that the purpose was limited? Will data be made public? What safeguards are in place for collection, use, and sharing?

- *Can subjects be identified?* What precautions are being taken to avoid this? The speaker noted that while some encryption is weak and can be broken, it still provides some protection.

- *What harms or risks have been identified for individuals or groups?* (In some cases, individual benefit may mean harm to group and vice versa.) Is there a potential for secondary use of data? How are data communicated? What is the extent of the risk? Are vulnerable populations involved? Is there possible harm to people who are not research subjects?

Triggers for more in-depth review include the use of personal identifiable information, new or unexpected uses of data, vulnerable populations, and an unfavorable benefit/risk analysis. The speaker stressed that the simple fact that the data are public should not mean the proposed project is a "done deal."

A model review process will provide an ethical review board with appropriate composition, accountability, adequate resources, transparent rules and procedures, safeguards, and formal reporting requirements or regulatory oversight. Review should reference ethical frameworks governing data use, address the proper role of consent (especially in an online context), and apply basic principles of fairness to affected individuals, groups and stakeholders.

Shared Responsibility in Ethical Big Data Research

- Jacob Metcalf, Ph.D.; Data & Society Research Institute

Dr. Metcalf used as a starting point a published article that presents "ten simple rules for responsible big data research" (Zook et al., 2017). He updated the rules, which were originally written for researchers, to suggest how they might apply to IRBs reviewing protocols for such research.

1. Acknowledge that data are people and can do harm.

"Start with the assumption that data are people (until proven otherwise), and use it to guide your analysis. No one gets an automatic pass on ethics."

IRB version: Start with the assumption that data are people even if the data scientists never "intervene" with those people. The Common Rule's exemption categories alone are not a satisfactory guide to ethical risk in data science.

2. Recognize that privacy is more than a binary value.

"Situate and contextualize your data to anticipate privacy breaches and minimize harm. The availability or perceived publicness of data does not guarantee lack of harm, nor does it mean that data creators consent to researchers using their data."

IRB version: When assessing privacy risk, ask researchers to address the extended ecosystem of privacy risks. Assume that data will be repurposed and recombined with third-party datasets. Consider the context of collection and ask researchers to adhere to that context, rather than binary categories of "public/private."

3. Guard against the reidentification of your data.

"Identify possible vectors of reidentification in your data. Work to minimize them in your published results to the greatest extent possible."

IRB Version: Require researchers to consider adversarial attempts to reidentify data of any released datasets.

4. Practice ethical data sharing.

"Share data as specified in research protocols, but proactively address concerns of potential harm from informally collected big data."

IRB Version: Require researchers to consider the downstream consequences of releasing models trained on sensitive human data, even if these data are already "public."

5. Consider the strengths and limitations of your data; big does not automatically mean better.

"Document the provenance and evolution of your data. Do not overstate clarity; acknowledge messiness and multiple meanings."

IRB Version: Consider whether big data studies, particularly those that involve sensitive classifications of people and/or vulnerable populations, meet the norms of beneficence and justice. Consider whether such studies may technically qualify as exempt, yet fail to satisfy core principles of the Belmont Report.

6. Debate the tough, ethical choices.

"Engage your colleagues and students about ethical practice for big data research."

IRB Version: Actively reach out to departments at your institution that are the hubs for data science but have not traditionally been considered sites of human subjects research (e.g., math, statistics, computer science/engineering). Offer training and guidance. Make sure they understand that "exempt" does not equal "responsible."

Dr. Metcalf stressed the importance of engaging researchers in ethical debate by reaching out to departments involved in data science. Most such researchers have not been trained in research ethics. If human research protection programs train them, these researchers can provide essential expertise on review panels.

7. Develop a code of conduct for your organization, research community, or industry.

"Establish appropriate codes of ethical conduct within your community. Make industry researchers and representatives of affected communities active contributors to this process."

IRB Version: Establish an internal strategy for addressing the risks specific to research involving data analytics, artificial intelligence (AI), and Machine Learning (ML). The norms are still evolving; contribute to that evolution by developing a local strategy and policies. Dr. Metcalf noted that Ms. Kasimatis Singleton has provided good examples from Johns Hopkins University School of Medicine.

8. Design your data and systems for auditability.

"Plan for and welcome audits of your big data practices."

IRB Version: Ask researchers handling sensitive human data how they will audit and ameliorate bias and fairness risks specific to data analytics and machine learning models, emphasizing their obligation to act consistently with the duties for beneficence and justice.

Dr. Metcalf noted that US lawmakers have introduced a bill, called the Algorithmic Accountability Act, that would require large companies to audit machine learning-powered systems — like facial recognition or ad targeting algorithms — for bias. If passed, the legislation would require the Federal Trade Commission to create rules for evaluating "highly sensitive" automated systems. Companies that have such systems would have to assess whether the algorithms that power them are biased or discriminatory, as well as whether they pose a privacy or security risk to consumers.

9. Engage with the broader consequences of data and analysis practices.

"Recognize that doing big data research has societal-wide effects."

IRB Version: Many, if not most, harms of data science research do not directly affect the subjects of the research. However, it is possible to address some downstream risks during collection, such as the treatment of mTurk workers, violations of ToS, and contextual integrity. Viral quizzes, such as those used in the Cambridge Analytica scandal, leverage weaknesses on the Internet and use deceptive practices to cause social impacts the data providers could not foresee.

10. Know when to break these rules.

"Understand that responsible big data research depends on more than meeting checklists."

IRB Version: Honestly, we probably could use more checklists.

Panel Discussion for Session III

Dr. Buchanan remarked that it has been frustrating to observe how the nature of this challenge keeps shifting over time. IRBs must keep retooling and relearning, trying to talk to computer scientists they are not used to dealing with. "Just when we think we know what we're doing, we're grappling with new stuff again."

What is "identifiable private information" today – and four years from now? Dr. Buchanan noted that the Common Rule stipulates that its current definition of identifiable private information will be examined every four years. What would conversations about this look like today, and what would you expect them to look like over time?

Dr. Zimmer hoped that the starting point for the discussion is that there is no such thing as nonidentifiable private information and we are checking in to see how we are addressing this problem. Dr. Garfinkel said this part of the law needs to be rewritten. The geoposition of a school bus reveals where a child with a cell phone went to school, and it is possible to get this information.

Ms. Kasimatis Singleton said the concept of identity being "readily ascertained" is more difficult, and IRBs need guidance on how to make that assessment.

What is "minimal risk" research in this environment? Observing that each day seems to bring news of new data breaches that expose people to risks, Dr. Buchanan wondered how the concept of "minimal risk" might be reevaluated as applied to big data research**.** How should IRBs assess whether such a project should be classified as "minimal risk"?

Ms. Kasimatis Singleton found this a challenging point. We are clearly facing more informational risk in our daily lives and need to consider whether proposed research is actually introducing a risk that is any different from those that already exist. She suggested that IRBs may need to focus on risk reduction by ensuring sufficient protections are in place to safeguard the data.

Further complicating the analysis, Dr. Garfinkel said, is the fact there easily could be three studies that are each minimal risk, but when all three are published they may, taken as a whole, place subjects at more significant risk.

Mr. Barnes suggested that IRBs have become more aware of the kinds of questions that must be asked before granting a waiver of consent. He offered the example of the Havasupai, a small identifiable group of people who offered their DNA for analysis related to a health issue that affected many tribe members and found their data had been passed on and used for additional purposes they would never have approved. Many IRBs waived consent for the research that arguably resulted in group harm to the Havasupai.

Are matrices sufficient to address security issues? Dr. Buchanan observed that matrices are emerging as a way of analyzing security risks. Are these sufficient from a philosophical standpoint? Dr. Zimmer responded that while they helpful as an operational tool, they can oversimplify issues and do not suffice.

How can data scientists be prepared to help and serve on IRBs? Dr. Buchanan asked panelists to comment on approaches to training data scientists to inform the deliberations of IRBs on big data projects. Dr. Zimmer suggested introducing ethics into data science curricula. Dr. Metcalf said more attention must be given to the kind of culture that is cultivated for data scientists: "How can we build a race to the top?" The best possible model should be offered to students.

Dr. Garfinkel noted that institutions and associations related to computer science have already developed codes of ethics, so it is really not about teaching them. There is really too much material to put in one single course. Dr. Bloss agreed that a single course on ethics isn't the optimum solution.

How do we evaluate the probability of risk or harm? Is new language needed? Dr. Buchanan wondered whether new language is needed when IRBs must review a protocol involving pervasive data and must assess the likelihood of risk or harm. Ms. Leong suggested that such an assessment must be made taking into account the likelihood of a particular harm occurring. If there is a high risk of a significant harm actually happening, then it probably shouldn't go forward. However, whatever measures are used to assess significant harm are inherently subjective and "depend on where you stand in relation to the harm."

How can ethics be introduced in unregulated spaces? Dr. Buchanan observed that much research, like the egregious political research that took place on Facebook in connection with Cambridge Analytica, is unregulated. How can ethics review occur in nonacademic environments? Dr. Metcalf responded that peers can review a proposed project and say it should not be allowed to go forward. Conference committees, conference papers, and journal articles, as well as statements of ethics, offer ways to gauge how peers view particular types of research.

Dr. Garfinkel opined that once the research is done, if it isn't published in one journal or presented at one conference, it may well be at another. However, many journals are now requiring preregistration, which is a helpful screening step.

Are there specific privacy concerns related to patient-reported clinical outcomes? A person listening to the webcast asked whether, with an increasing number of trials incorporating patient-reported clinical outcomes, there were any special concerns regarding the use of these data.

Ms. Kasimatis Singleton pointed to the technologies used to collect those outcomes. Some researchers want to use new technologies to extract data directly from individual health records or collect patient-reported outcomes via mobile apps. While IRBs once focused on what is being collected, they now also need to ask how information is collected. With ease of collection comes the need for more sophisticated analysis. Mr. Gupta agreed, stressing the importance of limiting the data researchers can access to those they actually need. Dr. Zimmer added that it is important for IRBs to review data management plans and determine how sensitive data are being stored.

Do we just need to "get over it"? Mr. Barnes said he had attended a conference in which a speaker from Sun Microsystems told those in attendance, "You have zero privacy. Get over it." Is there something we need to get over?

Dr. Garfinkel counseled that "anyone who says 'get over it' either feels invulnerable or wants to sell you something." These concerns arise in relation to protecting individuals and their families from unfair aspects of the world we live in. You don't say "get over it" if your child is being bullied. Even if you don't care about what happens to your own information, you should care about others who are vulnerable being able to make choices. Invasion of privacy is part of the core history of oppression.

Dr. Metcalf also said "get over it" is not the right approach. We do need to find the right amount of protection that will still allow us to improve the delivery of effective health care. Building the level of trust and protection for users of online systems is the next step. This means human labor is required to assess the context in which data are provided and used.

Final takeaways. In closing, Dr. Buchanan invited each panel member to highlight one takeaway from the workshop.

- *Dr. Kilpatrick:* We are dealing with a new form of systemic bias in the way we think about privacy, and we need to be cognizant about this as we think of individual controls.

- *Mr. Gupta:* We need to focus more on data protection and use as opposed to relying on informed consent for controls. "People don't get it."

- *Ms. Leong:* There is no single, final answer to this challenge. We will need to have an ongoing conversation, as with other social value-based systems. As technology changes, we will continue to adapt.

- *Dr. Zimmer:* There are many good examples of people trying to address the problem.

- *Dr. Li:* It is helpful to look toward a systems framework rather than relying on informed consent and individual-level controls.

- *Dr. Bloss*: We can focus on people as a point of intervention and help young kids, in particular, understand the issue. We need to remember that these data can do a lot of good.

- *Ms. Kasimatis Singleton:* IRBs need to think about this in terms of global impact of data use and develop a different framework for review. IRBs need to be prepared to look beyond the risk of a discrete project to understand the broader implications of data use.

- *Ms. Daniel:* Privacy and trust look different in different contexts. We need to consider not just the collection of data but the potential applications of models created on the basis of that data. We need to be concerned about the "life cycle" of the data collected and be prepared to apply different frameworks in different contexts.

- *Dr. Metcalf:* We have moved from initial panic and horror to a place where, though we are still horrified, we do have a sense of where to go next and we do have some tools we can get our hands on.

- *Dr. Garfinkel:* An abstract notion of privacy does not work for the types of analysis we are talking about. IRBs need to be able to look at social risks and benefits while also making sure results do not damage individuals. We need to move from open data toward trusted curation of public data.

- *Mr. Barnes:* We've highlighted the fact that so much of what is going on simply evades all the regulatory structures we have. At the end of the day we come back to the culture of ethics. Researchers can only use the data they can get, and we did not really talk about data governance in a deep way. Many institutions are carefully considering who should have access to data and for what purpose. The potential benefits of big data research are significant, but they sometimes come at the cost of individual privacy. For example, when New York City had a major epidemic of tuberculosis, mandatory surveillance was implemented that made it possible to ensure that individuals were taking their medications. This capacity was critical to ending the epidemic.

References

Buchanan, E., & Ess, C. (2009). Internet research ethics and the institutional review board: Current practices and issues. *ACM SIGCAS Computers and Society*, 39(3), 43-49.

Burt, A., Leong, B., Shirrell, S., & Wang, X. (2018). *Beyond explainability: A practical guide to managing risk in machine learning models* [White paper]. Retrieved September 19, 2019, from Future of Privacy Forum website: https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf.

Calo, R. (2013). Consumer subject review boards: A thought experiment. *Stanford Law Review Online*, *66*. Retrieved from https://review.law.stanford.edu/wp-content/uploads/sites/3/2016/08/Calo.pdf.

Chan, R, Jankovic F, Marinsek N, Foschini L, et al. (2019). *Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams*. Presented at SIGKDD Global Conference on Knowledge Discovery and Data Mining.

For patient-friendly description: https://www.kdd.org/kdd2019/accepted-papers/view/developing-measures-of-cognitive-impairment-in-the-real-world-from-consumer

Garfinkel, S.L. (April 14, 2008). *IRBs and Security Research: Myths, Facts and Mission Creep.* Presented at Usability, Psychology, and Security Conference, San Francisco.

Garfinkel, S.L., & Cranor, L.F. (2010). Institutional Review Boards and your research. *Communications of the ACM*, *53*(6), 38-40.

McDonald, A.M., & Cranor, L.F. (2008). The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, *4*(3), 543-568.

Polonetsky, J., Tene, O., & Jerome, J. (2015). Beyond the common rule: Ethical structures for data research in non-academic settings. *Colorado Technology Law Journal*, *13*, 333-367.

Privacy Nutrition Labels. (n.d.). Retrieved from https://cups.cs.cmu.edu/privacyLabel/.

Schroepfer, M. (2014, October 2). Research at Facebook [Blog post]. Retrieved from http://newsroom.fb.com/news/2014/10/research-at-facebook/.

Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics*, *12*(5), 372-382.

Zook, M., Barocas, S., Boyd D., Crawford, K., Keller, E., Gangadharan, et al. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology*, *13*(3). Retrieved from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399

Online Resources

Resources for Centers for Medicare & Medicaid Services (CMS) data are listed below:

- Research Data Assistance Center website
- List of available data at the Research Identifiable File Availability page of the Research Data Assistance Center
- The Chronic Conditions Data Warehouse website
- The Chronic Conditions Data Warehouse's list of Frequently Asked Questions

- Public use files are available at CMS' Research, Statistics, Data & Systems page

Resources for codes of ethics are listed below:

- The Association for Computing Machinery's Code of Ethics and Professional Conduct
- The Data Science Association's Code of Conduct: Data Science Code of Professional Conduct
- The Department of Homeland Security's Fair Information Practice Principles

Resources for Privacy and Health Research in a Data-Driven World are listed below:

- Recorded webcast
- Presentation slides
- Additional information and links for the workshop

Other relevant resources and organizations are listed below:

- The Council for Big Data, Ethics, and Society website
- The Future of Privacy Forum website
- The International Association of Privacy Professionals website
- The Pervasive Data Ethics for Computational Research (PERVADE) website
- U.S. Census Bureau, Selected Resources on Disclosure Avoidance:
    - Presentation slides on Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test (Continued) from a presentation given on September 15, 2017 at the 2017 Census Scientific Advisory Committee Fall Meeting
    - A blog from the U.S. Census Bureau from June 6, 2019, entitled Disclosure Avoidance and the 2018 Census Test: Release of the Source Code. This blog explains how to use the code base with the 1940 Census public data from IPUMS.
    - John M. Abowd, the Chief Scientist and Associate Director for Research and Methodology at the U.S. Census Bureau, gave the keynote address at the International Conference on Machine Learning on June 11, 2019. His talk was titled "The U.S. Census Bureau Tries to Be a Good Data Steward in the 21st Century."
        - Read the abstract of his talk.
        - Watch the video of his talk.
    - The U.S. Census Bureau's Memorandum 2019.13: Disclosure Avoidance System Design Parameters and Global Privacy-Loss Budget for the 2018 End-to-End Census Test
- The Vivli Center for Global Clinical Research Data website