HIT Policy Committee
Privacy and Security Workgroup
FINAL Report of the December 5, 2014 Virtual Hearing on Health Big Data

Name of ONC Staff Liaison Present: Kathryn Marchesini

Meeting Attendance: (see below)

Purpose of Hearing:

Workgroup Chairperson Deven McGraw reported that according to a May 2014 report from the Executive Office of the President, "The government should lead a consultative process to assess how the Health Insurance Portability and Accountability Act (HIPAA) and other relevant federal laws and regulations can best accommodate the advances in medical science and cost reduction in health care delivery enabled by big data." The charge is to ensure privacy protection for big data analyses in health. Big data introduces new opportunities to advance medicine and science, improve health care, and support better public health. To ensure that individual privacy is protected while capitalizing on new technologies and data, the Administration, led by the Department of Health and Human Services, will: consult with stakeholders to assess how federal laws and regulations can best accommodate big data analyses that promise to advance medical science and reduce health care costs; and develop recommendations for ways to promote and facilitate research through access to data while safeguarding patient privacy and autonomy. McGraw stated the purpose of the hearing:

- Listen to and learn from leading experts with diverse perspectives on issues related to big data in health care
- Engage in robust discussions to better understand the landscape, opportunities, and challenges
- Construct a base of knowledge that will be further augmented over the coming months as the work group develops recommendations for the HITPC and ultimately ONC

Each invited presenter was allowed 5 minutes. They were invited to submit written testimony as well.

Health Big Data Opportunities and the Learning Health System

Social determinants of health and personally generated data: Steve Downs, Robert Wood Johnson Foundation, talked about a culture of health and the opportunity to understand more about the impact of social determinants. One can imagine health care systems leveraging data on housing stock, on community walkability, safety and violence, availability of early childhood services, food accessibility, transportation infrastructure and more to understand the barriers faced by individual patients and by the population as a whole. For instance, Ruben Amarasingham, at Parkland Health and Hospital System in Dallas, has been using data on insurance status and the number of home addresses reported in the past year to develop predictive algorithms to understand which heart failure patients are at risk for readmission. By providing additional services to those determined to be at greatest risk, Parkland reduced readmission rates by 30%. Apps are starting to provide a window onto people's day-to-day experience with health. Other sources such as supermarket loyalty card data or credit card purchase data could fill out the picture even further. These data could be used for research purposes and also for the practice of public health. Three characteristics of personally generated data -- the breadth of

variables about day-to-day experience that can now be captured, the near continuous nature of its collection and the sheer numbers of people generating the data -- make it extremely interesting for research. He imagined research that explores the relationships between neighborhood walkability and steps taken, between food access and dietary patterns, or between stress and geographic location. RWJF projects identified issues to be overcome: privacy, informed consent, access to the data and data quality. The skewed adoption of smartphone apps and wearable sensors poses methodological challenges. RWJF has funded a network of researchers and others to work on these issues. He described examples in which PGHD may be applicable to public health questions. He called for experimentation, for the technology and the methods to get better, and to allow our institutions to catch up so that they can learn how best to take advantage of these opportunities.

Experience of the FDA Mini-Sentinel program: Rich Platt, Harvard Pilgrim HealthCare Institute, acknowledged that most clinical guidelines are based on limited information. Better data are needed. He said that EHRs and billing data are irreplaceable for answering many clinical questions, such as effectiveness and safety of medical practices, understanding which treatments work best for specific groups, quality of care, overall and in specific health systems, assessing health status of communities, and guiding public health interventions and measuring impact. EHRs can also be used in identifying patients who might want to participate in clinical trials. But some questions require use of fully identified information, e.g., linking to the National Death Index. It is not possible to obtain individual consent for all uses of individuals' data. It is not possible to notify all individuals personally about all uses of their data. Opt-out provisions can make answers to questions unreliable. He said that the minimum necessary amount of identifiable data should be used. Approval and oversight should always be required. The specific uses of data should be stated publicly and the number of individuals with access to personal medical information should be minimized. He showed slides that described the use of the Mini-Sentinel Distributed Database, which contains 358 million person-years of observation time. He provided examples of its use. Distributed data analysis can eliminate or greatly reduce the need to transfer personally identifiable data. Personally identifiable information is required in some circumstances, e.g., linking an individual's data across two sources. Identifiable information should be stored in protected locations. Data enclaves are one solution. Notifying patients of data sharing practice at the time of care as well as public notice to the community via multiple means is important.

Incorporating patient data and learning from it: Patti Brennan, University of Wisconsin-Madison College of Engineering, referred to tiny, medium sized, and big data. People are experts in every-day living. Incorporating patient-defined data into the health data flow is important. Information must flow two ways over time. She delineated requirements for a robust data infrastructure, such as robust network connections, access control, privacy mechanisms, and interoperability. Both local and large-scale analyses are important. She referred to a report—Capturing Social and Behavioral Domains in Electronic Health Records—available at: www.iom.edu.

Q and A

Downs observed that user demographic characteristics are dynamic and changing. Caution should be taken in drawing conclusions. But because of the extraordinarily large Ns, there is some support for a representative sample. Brennan pointed out that in the future environmentally embedded sensors will generate data, for example, temperatures gathered among airport passengers, motion sensors, noise as an indicator of crowd density and air quality.

In response to another question, Downs referred to the citations in his written testimony. He said that the vast majority of individuals would allow their health data to be used for research purposes if identity were protected. Consent is nuanced.

David McCallie inquired about unintended consequences of these data and how to mitigate them. He referred to red lining and decisions on where to invest. Panelists mentioned adverse selection of patients and subscribers, inadvertent disclosure of illegal activity, and valuing of objective over subjective data, such as judgments and feelings. Most likely, there are many potential unintended consequences. There must be sanctions for misuse.

When asked to give an example of the need for identifiable data for research, Platt explained that data on patients' identity is needed for linking records, for example, linking state influenza immunization registry data with EHRs, since vaccines are often administered in a variety of sites, to study complications resulting from vaccinations. Another example is linking clinical records to the National Death Index for research on outcomes of clinical interventions.

McCallie wondered about the threat of sanctions reducing the need for redactions. According to Platt, both are needed. He emphasized that much can be learned from small amounts of data transferred. Brenman observed that protections are in place in the clinical setting, but much other data lack any formal protection. Downs pointed out that one could easily conclude that all data are related to health. With the increasing ability to identify individuals, redactions may be easy to circumvent.

Linda Kloss referred to covered entities (CE) and wondered about designating covered data, which are not tied to the holder. Downs replied that it would be difficult to draw a line for what to cover. For example, even credit scores can be used to infer information about medication adherence. Individuals have different concerns. Combinations of data can be used to infer. Platt emphasized the importance of focusing on the use of data and oversight.

McGraw asked about harmless use. How is it defined? Is commercial use harmful? Platt indicated that any discrimination or denial of opportunity or public embarrassment would obviously be harmful, as well as commercial use without transparency as to its use. McGraw went on to inquire about creepiness. Brenman talked about a study of phone call length and inferences regarding social behavior conducted in a small community, saying that knowledge of one's friendship network could lead to social sanctions. People fear how government may use their data in the future. Platt said that public disclosure of the use of data and oversight should provide confidence. Downs pointed out the fine line between carefully tailored services and creepiness.

Regarding processes developed in academia that can be used to control the Mosaic effect, Brenman indicated that discourse on effects can be helpful to be sure proposers are aware of both intended and unintended consequences. Data safety monitoring plans can be expanded. Journals are now asking researchers to deposit the data used for their research papers. Downs said that it is possible to place controls on the data and the users. Controls can be a part of the approval process. Workgroup Co-chairperson Stan Crosley commented that it will be difficult when data come from several different places. Brenman talked about the issue of a user gaining financial gain without sharing gain with contributors. Public disclosure is important. Others talked about allowing individual control. But if individuals can opt-out, the representativeness of the sample is lost. There should probably be things that are always prohibited. An ONC staff member raised the idea of a social contract.

Health Big Data Concerns

Health privacy and security: Michele DeMooy, Center for Democracy and Technology, said that many of the challenges facing traditional health care providers in the big data era also apply to app developers and wearable device manufacturers. Notice and consent remains a problem, especially given the decreased ability to read notices on mobile device screens or via a wearable device. HIPAA does not apply to most app developers or device manufacturers. Developers and device manufacturers should consider incorporating privacy and security protective measures, based on the FIPPs, into their products. The Privacy Act of 1974 was designed to give individuals some control over personally identifiable information collected about them by the federal agencies. The law applies to any federal agency that provides health care services for the government, as well as agency contractors that are considered HIPAA-covered entities. There are exceptions for disclosure for administrative uses and public health and safety emergencies. When information is de-identified, government entities do not need patient consent to collect and use it and it is not covered by the Privacy Act; however, many federal and state agencies choose to guide privacy and disclosure for de-identified data on ethical guidelines that review the implications of revealing data, regardless of law or policy. U.S. citizens and permanent residents have a right to access, inspect and potentially amend health records maintained by the government. One concern with current legal regulation of government use of health data is that non-citizens' only means of access and amendment is HIPAA, which arguably provides less privacy protections that the Privacy Act. FIPP offers governments seeking to use big data with regard to health information a strong, standardized structure that promotes responsible and efficient use of data while allowing for innovations in analytics and application. Transparency should guide any health data collection and use regime, from the first point of contact with data to any subsequent use. Information about data practices can be done in different ways. Privacy policies tend to be inscrutable, risk-averse compliance obligations, in which the primary goal is to avoid making an incorrect statement that could serve as the basis for FTC liability. It is particularly important to get notice right when using data collection methods that are less visible, such as collection from mobile health applications that typically involve individuals inputting their own health data. To mitigate the opacity of this collection, entities are obligated to make full disclosure to those they collect from about data practices via contextual notice. Contextual notice, or just-in-time notice, is a critical component of meeting an individual's collection and sharing expectations. Fundamentally, it should be clear to a consumer using a health app or wearable device when data are being collected, what types of data are being collected, what they are used for, what secondary uses of the data are contemplated, how long data are retained, and what security measures are put into place in order to protect the data.

e-Health: Mark Savage, National Partnership for Women and Families, reported on a recent conference on big data and civil rights hosted by the Data and Society Research Institute, Leadership Conference on Civil and Human Rights and New America's Open Technology Institute, on October 30, 2014. Two distinct issues were at the forefront. One was the threat that greater surveillance (big data collection and analysis) poses for low-income communities and communities of color. The other issue is health disparities, and the promise of big data in identifying, analyzing and addressing health disparities. He mentioned themes from the conference. The same piece of data can be used to reduce health disparities and empower people, or conversely, to violate privacy and cause harm—depending on who holds the data and what the person does with the data. Greater demographic granularity can help to address health disparities or increase the risk of profiling. For instance, information about whether one has had a vaccination could be used for a public education campaign, or for targeting an increase in insurance premiums. Additionally, all data can be health data, or data from which inferences about health are drawn or correlations with health are made. He said that the focus should be on uses and

harms rather than costs and benefits. Discussing costs and benefits implies trade-offs, and people thought it premature to focus on that calculation. Focusing on harms helps to seek redress through civil rights laws. Savage went on to report on another conference that eventually resulted in nine patient-consumer principles: benefits for personal health and population health; ensuring that all patients and consumers benefit fully and equally; designing the technology and services to meet the range of needs without barriers or diminished function for some communities; ensuring the privacy and security of patients' health information; preventing misuse of patients' data; building partnership and HIT literacy among patients, providers, and public health workers; and accountability for achieving the benefits of health information exchange.

Patient perspective: Anna McCollister-Slipp, Galileo Analytics, gave a graphic description of her personal experience as a patient. She delineated some of the reasons given for limiting individuals' access to their own data such as the desire to monetize the data. Patients are told: the EHR will not allow it; the data must be protected; or HIPAA prevents it. Patients are frustrated because the volume of information is increasing but patients cannot use it. Perceptions of security and privacy are barriers. There is an urgent need for data liquidity. People who have the data will not share it. There is no sense of urgency about sharing data.

Q and A

In response to a question about algorithm transparency and its similarity to a learning system, DeMooy explained that algorithms take information and contextualize it. To some extent they are replacing human decision making. Therefore, it is important to look at the demographic characteristics of their designers since those individuals eventually affect decisions. Algorithms are considered proprietary, but they should be transparent regarding the variables used.

Linda Kloss wondered about the lack of understanding of privacy being a barrier. DeMooy argued that transparency and more information about what happens would eventually contribute to trust. People misunderstand the range of HIPAA. McCollister-Slipp argued that misunderstanding of HIPAA can be an excuse for lack of progress. According to Savage, consumers want both protection and the use of their information. There are tools to do both. McCollister-Slipp noted that there is confusion even in government agencies about the interpretation of privacy.

McGraw asked which privacy procedures unnecessarily create obstacles. McCollister-Slipp described her own unsuccessful efforts to obtain her data. She referred to a Pew survey and threat models for communities and individuals. Savage opined that existing laws provide an adequate framework for protection, but there should be a prohibition against re-identifying data.

McCollister-Slipp responded to a question about whether any of her efforts actually worked by saying that recently her doctors have been freer will e-mail communication, mostly using their personal accounts. She continues to be unable to obtain labs results directly in the format she prefers.

According to one of the presenter, civil rights law provides a good framework. Some uses of data are always permissible and others should always be prohibited. McCollister-Slipp opined that when data are de-identified, patients should not have the right to withhold their use. The data should be available to researchers who are working for the good of patients. DeMooy agreed, also pointing out that some data can be re-identified. Research on anonymization is needed.

McGraw asked about customary restrictions on data collection. DeMooy seemed to agree that not all identifiable items are required for the data to talk. McCollister-Slipp repeated her opinion of de-

identified data. She declined to comment on the level of de-identification, saying that she is not an expert.

In response to a comment about the prevention of unintended harm and the difficulty of proving inadvertent impact, DeMooy indicated that some factors are surfacing, such as who makes the decisions in the design of algorithms. Savage referred to the FIPP collection limitations, which may not be applied as well as they could be. DeMooy suggested exploration into diagnostic areas. McCollister-Slipp pointed out that the data on which clinical guidelines are based are not necessarily representative in that minorities are under-represented in clinical trials. McCallie referred to the prohibition of denial of insurance due to preexisting conditions as an example of policy intended to mitigate harm.

Protections for Consumers

De-identification and encryption: Khaled El Emam, **University of Ottawa**, made six key points. The narrative on re-identification attacks is misleading and not based on an accurate interpretation of the evidence, leading data custodians to stop sharing data. A better informed conversation and a proper evidence-based narrative about re-identification risks are needed. There is a need for de-identification standards to help with the adoption of good practices and make it easier for data custodians to share data responsibly. The risks from continued use of the HIPAA Privacy Rule Safe Harbor de-identification method need to be evaluated. There is a need for ethics councils within organizations that perform analytics on health data to manage the risks from inferences, and guidance is required on how to set up and run such councils. This is a practical way to manage inference risks from data. The main benefit from safe havens is that they would allow the sharing of data sets that have experienced less de-identification than when raw data are shared. But that data still need to be de-identified to some extent. Before deploying distributed computation systems in practice, it is important to develop more formal security proofs for them. A number of systems deployed today may be leaking personal information under certain circumstances. He went on to say that there are generally two types of disclosure that are considered in the context of disclosure control that need to be protected against. One type of disclosure is identity disclosure, which is when an adversary is able to assign a correct identity to a record. When discussing identifiability of non-personal information, one is referring to that type of disclosure – that of assigning an identity to a record. A de-identified data set is one where the probability of identity disclosure is very small. One must distinguish between different types of health data. Genomic data is difficult to de-identify and good methods for doing so are still in the research phase. These methods utilize secure computation protocols. It will likely be a few years before this research produces results that can be scaled into practice. Other types of data, such as administrative, clinical, trials, and survey data, can be de-identified using existing techniques. There are good de-identification techniques that are based on generally accepted statistical and scientific principles. All known re-identification attacks that have been successful on this type of data were performed on data sets that were not de-identified properly. While the HIPAA Privacy Rule provides some methods for de-identification, these need to be operationalized into detailed standards that individuals can follow. The lack of standards means that data custodians have to figure out what they need to do on their own. There is not a sufficient pool of de-identification experts and an ecosystem of services to support de-identification. Regarding the use of data to make decisions that confer stigma, governance should be used to manage the use. Data access committees can be formed. Safe havens have both advantages and disadvantages. The data going to safe havens must be defined and there is still some risk of re-identification.

The need for patient control and fair information practices: Bob Gellman, **Private Consultant**, pointed out the absence of a precise definition of big data. A definition must precede regulation. People who have always opposed the Privacy Rule have jumped on the big data concept as a way to avoid

restrictions. Big data advocates are making unrealistic promises. The Google flu trend did not work. There is already a mechanism for making patient data available for research. HIPAA provides a common set of rules, which is good.

Limits of consent and appropriate data use: Fred Cate, Indiana University Maurer School of Law, said that privacy law seems overly focused on the individual. The definition of privacy is inadequate and sets up an impossible expectation. Control of one's own data has many disadvantages and risks. Patients and caregivers want their data to be used for research to improve health conditions. However, according to his research findings, they do not want to be contacted and bothered, even for consent. They wish to avoid interruptions. He said that privacy is too important to leave up to individual policing. The demand for data is increasing, and not only for big data. Protections should not interfere with benefits.

Q and A

McCallie questioned El Emam on re-identification risks when data sets are jointed. El Eman repeated that there is no evidence that this occurs. It is hard to do attacks, because when data are shared, there are additional controls. Contracts can prohibit joining data with other sets. There are ways to modify data to add protection. Controls plus techniques for joining sets provide protections. McGraw agreed to circulate Gellman's article on a suggested legislative framework.

Regarding safe havens, El Emam said that enclaves are the same as havens; they create a closed environment. But safe harbors are based on assumptions that are not always met. Expert methods often lack transparency, or there is no oversight as to the qualifications of the expert. He seemed to agree that licensure standards for oversight or certification for experts would be a good idea. A greater pool of experts is needed. Gellman said that publication of the methodology so that it could be independently analyzed would be good.

McGraw observed that the safe harbor categories of information have not been examined since the law was enacted; it would be good to have a simply applied method. El Emam indicated that stronger methods should be more adoptable. Gellman referred to data use agreements, saying that limited information can be made public. Some things should be discussed in public.

Regarding harms, HIPAA addresses identification through a process. Requirements against re-identification should be strengthened. Stronger legal requirements are needed. But privacy protections should not create new risks. Appropriate uses for research should be clarified. Gellman urged them to consider rights, not just potential harms. The compilation of data in itself may be harmful. Cate and Gellman disagreed not only on the individual's rights to control data, but also on what individuals say they want. They referred to various polls and reports to support their conflicting viewpoints.

Considering consent, Cate indicated that there must be meaningful things to which to consent. Consent is used as an opt-out. Consent should focus on things people actually care about, such as sharing their data with the government, which is never an option. Gellman said that the purpose of IRBs has been to allow society to consent for the greater good. But the technology is now available to find people and get their consent directly.

In response to a question about standards, El Emam said that some standards are being developed for the de-identification of general health data. Some standards are also available for clinical trials and there are policies around data sharing standards in development. Other industry efforts are expected for release next year. Standards should be data or use specific. But they must be operational, automated, and scalable. Someone observed that IRBs want guidance because they often get conflicting advice. Cate said that guidance should be based on the use of the data. His experience with large trials is that

different IRBs give conflicting advice. Cate and Gellman agreed that much, possibly most, health data are not protected by HIPAA.

McGraw said that the earlier presenters recommended more transparency on the actual use of data. Someone agreed that more transparency regarding IRBs is needed. IRB reviews should be made public. Another presenter agreed that notices should not be confused with transparency. Notices are meaningless. Transparency is an ethical concern.

The hearing continued on December 8.

Public Comment: None

Meeting materials:

Presentation slides
Written testimonies
Questions
Bios
Background submissions

Meeting Attendance			
Name	**12/05/14**	**11/24/14**	**11/10/14**
Adrienne Ficchi			
Bakul Patel			
Cora Tung Han	X		
David Kotz		X	X
David McCallie, Jr.	X	X	X
Deb Bass			
Deven McGraw	X	X	X
Donna Cryer	X	X	X
Gayle B. Harrell	X	X	X
Gilad Kuperman			X
Gwynne L. Jenkins			
Helen Caton-Peters	X		X
John Wilbanks			
Kathryn Marchesini	X	X	X
Kitt Winter	X	X	X

Kristen Anderson	X	X	X
Linda Kloss	X	X	X
Linda Sanches	X	X	X
Manuj Lal			
Mark Sugrue			X
Micky Tripathi	X	X	
Stanley Crosley	X	X	X
Stephania Griffin	X		
Taha A. Kass-Hout	X	X	
Total Attendees	**15**	**13**	**14**