Thursday May 26, 2016
Via electronic filing

Linda L. Kloss, MA
Chair, NCVHS Subcommittee on Privacy, Confidentiality and Security
President, Strategic Advisors, Ltd.
PO Box 290
Sister Bay, WI 54234
Email: linda@kloss-strategicadvisors.com

Re: *Hearing, Subcommittee on Privacy, Confidentiality & Security; National Committee on Vital and Health Statistics*

Dear Ms. Kloss and Members of the Subcommittee,

This comment is informed by research with collaborators through the *Privacy Tools for Sharing Research Data* project at Harvard University.[1] In this broad, multidisciplinary project, we are exploring the privacy issues that arise when collecting, analyzing, and disseminating research datasets containing personal information. Our efforts are focused on translating the theoretical promise of new measures for privacy protection and data utility into practical tools and approaches. In particular, our work aims to help realize the tremendous potential from social science research data by making it easier for researchers to share their data using privacy protective tools.

Academic research in theoretical computer science, statistics and information science has demonstrated a number of challenges related to managing information privacy in the modern world. For example, a recent workshop exploring challenges related to privacy and statistical data sharing illustrated three challenges, paraphrased below:[2]

The first challenge is that many human behaviors leave behind distinct behavioral fingerprints in the data, even in the complete absence of pieces of information considered to be identifiers (or quasi-identifiers). This creates a problem for most traditional statistical disclosure limitation methods that aim to suppress or perturb the identifiers in a dataset. A second challenge is that when data are released after traditional statistical disclosure control methods such as identifier-based redaction or aggregation to prevent record-linkage have been applied, informational risks to individuals from that data release continue to grow in the future as new external data are released. This is because traditional methods, regardless of the modification they make to the data (e.g. swapping, top or bottom coding, generalization, noise addition, etc.), do not address the accumulation of risk from multiple releases of data. While these methods may in some cases be sufficient for controlling what can be learned about an individual from a

[1] The Privacy Tools for Sharing Research Data project is supported by a National Science Foundation Secure and Trustworthy Cyberspace Frontier grant and a grant from the Alfred P. Sloan Foundation. See Privacy Tools for Sharing Research Data, http://privacytools.seas.harvard.edu.

[2] Altman M, Capps C, Prevost R. Location Confidentiality and Official Surveys. Social Science Research Network [Internet]. 2016.

specific data set, modern privacy research shows that such approaches cannot provide any strict bounds on the amount that can be learned through composition with independent auxiliary information. A third challenge revealed by modern privacy research is that every release of data, if it has any utility, no matter how it is protected, inevitably leaks some private information, and this leakage increases with each release of data. In other words, there is no free lunch with respect to information privacy; you always have to buy it with utility.

In previous publications and regulatory comments, my collaborators and I have offered a number of recommendations that we believe would help enable the wider sharing of research data while providing privacy protection for individuals.[3]

Although our previous writings do not comment directly on the HIPAA regulations and safeguards, it is my judgement that the risks discussed in these works apply to protected health information, and that the broad findings and recommendations are readily applicable to improving HIPAA. For these reasons, I recommend that the committee read and incorporate these recommendations, which are summarized below.

As a general framework, my collaborators and I have recommended the development of rules and guidance based on the following principles of a modern approach to privacy:[4]

- Calibrating privacy and security controls to the intended uses and privacy risks associated with the data;
- When conceptualizing informational risks, considering not just reidentification risks but also inference risks, or the potential for others to learn about individuals from the inclusion of their information in the data;
- Addressing informational risks using a combination of privacy and security controls rather than relying on a single control such as consent or deidentification;
- Anticipating, regulating, monitoring, and reviewing interactions with data across all stages of the lifecycle (including the post-access stages), as risks and methods will evolve over time; and
- In efforts to harmonize approaches across regulations and institutional policies, emphasizing the need to provide similar levels of protection to research activities that pose similar risks.

(We note in prior writings that terms above, such as privacy, confidentiality, security, and

[3] See Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Technology Law Journal 2015;30(3):1967-2072; Wood A, Airoldi E, Altman M, de Montjoye Y-A, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rulemaking. Comments on Regulations.gov. 2016. (Copy available here: http://informatics.mit.edu/publications/privacy-tools-project-response-common-rule-notice-proposed-rule-making); Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Elements of a New Ethical and Regulatory Framework for Big Data Research. Washington and Lee Law Review. 2016;72(3):420-442.

[4] Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Technology Law Journal. 2015;30(3):1967-2072; Wood A, Airoldi E, Altman M, de Montjoye Y-A, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rulemaking. Comments on Regulations.gov. 2016.

sensitivity are used in multiple communities of practice in somewhat different ways, and they are defined inconsistently throughout the literature. We suggest a vocabulary for these terms in the works cited above, and I recommend that any regulation refer to explicit definitions of these terms.)

In related work, we have argued for the need for comprehensive and consistent regulatory protection against information privacy harms in research. Protection for people whose information is used in research should be based on the risks and benefits to the subject and to society, and not on other elements of the research context that are irrelevant from an ethical perspective, such as the institution conducting the research, its commercial status, or its sources of funding.[5]

In addition, the research cited above finds that addressing privacy risks requires a sophisticated approach, and the privacy protections currently employed in government releases of data do not take into account recent advances in data privacy research. We note that there is a wide range of technical, procedural, legal, educational, and economic controls available for managing privacy risks. However, most government data releases rely almost exclusively on a narrow set of interventions, namely redaction of identifiers and binary access control. This focus on a small set of controls likely fails to address the nuances of data privacy and utility, as well as the differences between data releases, which vary widely in terms of the intended uses of the data and the privacy risks involved.

This research also notes, as paraphrased, that advances in science and technology enable the increasingly sophisticated characterization of privacy risks and harms and offer new interventions for protecting data subjects. In our work,[6] we describe a lifecycle approach that supports a systematic decomposition of the factors relevant to data management at each information stage, including the collection, transformation, retention, access or release, and post-access stages. Additionally, we propose a framework for developing guidance on selecting appropriate privacy and security measures that are calibrated to the context, intended uses, threats, harms, and vulnerabilities associated with a specific research activity.

Figure 1 provides a partial conceptualization of this framework.[7] In this diagram, the x-axis provides a scale for the level of expected harm from uncontrolled use of the data, meaning the maximum harm the release could cause to some individual in the data based on the sensitivity of the information. This scale ranges from low to high levels of expected harm, with harm defined to capture the magnitude and duration of the impact a misuse of the data would have on an affected individual's life, and we have placed examples as reference points along this axis. The y-axis provides a scale for the post-transformation identifiability, or the potential for others to

[5] Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Elements of a New Ethical and Regulatory Framework for Big Data Research. Washington and Lee Law Review Online. 2016;72(3):420-442.

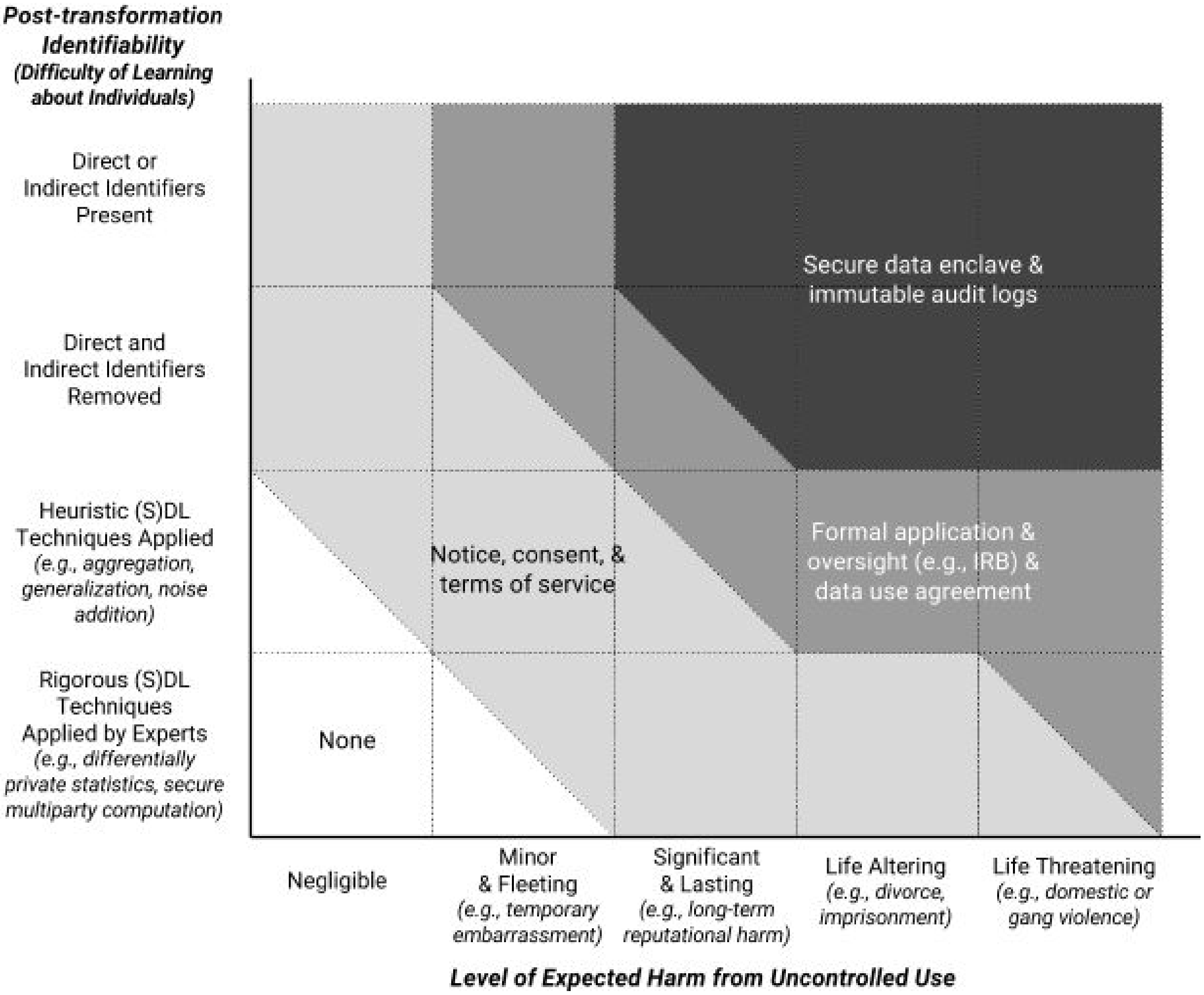
[6] Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Technology Law Journal. 2015; 30(3): 1967-2072.

[7] This diagram originally appeared in Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Technology Law Journal. 2015; 30(3): 1967-2072.

learn about individuals based on the inclusion of their information in the data. A number of examples are provided as anchor points, ranging from data sets containing direct or indirect identifiers, to data shared using expertly applied rigorous disclosure limitation techniques backed by a formal mathematical proof of privacy.

The level of expected harm from uncontrolled use and the post-transformation identifiability of the data, taken together, point to minimum privacy and security controls that are appropriate in a given case, as shown by the shaded regions in the diagram. Regions divided by a diagonal line correspond to categories of information for which an actor could reach different conclusions based on the intended uses of the data or privacy standards that vary based on the applicability of a regulation, contract, institutional policy, or best practice. The sets of controls within the shaded regions focus on a subset of controls from the more comprehensive set of procedural, economic, educational, legal, and technical controls we catalog below in Table 1. In practice, the design of a data management plan should draw from the wide range of available interventions and incorporate controls at each stage of the lifecycle, including the post-access stage. Also note there are regions of this diagram that deviate from current practice in some domains. For example, we argue that data that have been de-identified using simple redaction or other heuristic techniques should in many cases be protected using additional controls, even though some existing standards do not expressly call for the use of additional controls when using such techniques.

Figure 1. Calibrating privacy and security controls.


Post-transformation Identifiability
(Difficulty of Learning about Individuals)
Direct or Indirect Identifiers Present
Direct and Indirect Identifiers Removed
Heuristic (S)DL Techniques Applied
(e.g., aggregation, generalization, noise addition)
Rigorous (S)DL Techniques Applied by Experts
(e.g., differentially private statistics, secure multiparty computation)
Secure data enclave & immutable audit logs
Notice, consent, & terms of service
Formal application & oversight (e.g., IRB) & data use agreement
None
Negligible
Minor & Fleeting
(e.g., temporary embarrassment)
Significant & Lasting
(e.g., long-term reputational harm)
Life Altering
(e.g., divorce, imprisonment)
Life Threatening
(e.g., domestic or gang violence)
Level of Expected Harm from Uncontrolled Use

For many activities, implementing a single set of privacy and security controls may not be appropriate for all intended uses of the information. For this reason, we generally recommend that regulators and data controllers implement a tiered access model. A tiered access model is one in which data are made available to different categories of data users through different mechanisms.

Figure 1 illustrates the relationship between transformation and release controls, and suggests how controls could be selected for different access tiers. For example, an investigator could provide public access to some data without restriction after robust disclosure limitation techniques have transformed the data into differentially private statistics. Data users who intend to perform analyses that require the full dataset, including direct and indirect identifiers, could be instructed to submit an application to an oversight body such as an institutional review board, and their use of the data would be restricted by the terms of a data use agreement. We argue that this framework, implemented through a data management plan and tiered access model, would help data providers, data users, and oversight bodies calibrate the use of privacy and security controls to the contexts, threats, harms, and vulnerabilities associated with a research activity, as well as the purposes desired by different categories of data users.

Table 1 below provides an example catalog illustrating the wide range of procedural, economic, educational, legal, and technical controls that are available at each lifecycle stage. Data providers should be encouraged to consider incorporating within their data management plans a combination of privacy and security controls drawn from the wide range of available of interventions, rather than relying on a single control such as redaction or binary access control at the release stage.

Table 1. Example catalog of privacy and security controls.

	Procedural	Economic	Educational	Legal	Technical

Collection/ Acceptance	Collection limitation;
	Data minimization;
	Data protection officer;
	Institutional review boards;
	Notice and consent procedures;
	Purpose specification;
	Privacy impact assessments
		Collection fees;
		Markets for personal data;
		Property rights assignment
			Consent education;
			Transparency;
			Notice;
			Nutrition labels;
			Public education;
			Privacy icons
				Data minimization;
				Notice and consent;
				Purpose specification
					Computable policy

Transformation	Process for correction		Metadata;
			Transparency
				Right to correct or amend;
				Safe harbor de-identificati on standards
					Aggregate statistics;
					Computable policy;
					Contingency tables;
					Data visualizations;
					Differentially private data summaries;
					Redaction;
					SDL techniques;
					Synthetic data

Retention	Audits;
	Controlled backups;
	Purpose specification;
	Security assessments;
	Tethering
			Data asset registers;
			Notice;
			Transparency
				Breach reporting requirements;
				Data retention and destruction requirements;
				Integrity and accuracy requirements
					Computable policy;
					Encryption;
					Key management (and Secret sharing);
					Federated databases;
					Personal data stores

Access/Release	Access controls;
	Consent;
	Expert panels;
	Individual privacy settings;
	Presumption of openness vs. privacy;
	Purpose specification;
	Registration;
	Restrictions on use by data controller;
	Risk assessments
		Access/Use Fees (for data controller or subjects);
		Property rights assignment
			Data asset registers;
			Notice;
			Transparency
				Integrity and accuracy requirements;
				Data use agreements (contract with data recipient)/ Terms of service
					Authentication;
					Computable policy;
					Differential privacy;
					Encryption (incl. Functional;
					Homomorphic);
					Interactive query systems;
					Secure multiparty computation

Post-Access (Audit, Review)	Audit procedures;
	Ethical codes;
	Tethering
		Fines	Privacy dashboard;
			Transparency
				Civil and criminal penalties;
				Data use agreements/ Terms of service;
				Private right of action
					Computable policy;
					Immutable audit logs;
					Personal data stores

In our prior work,[8] we also call special attention to advanced data-sharing models and emerging formal approaches to privacy. We note that there are a number of privacy methods and data-sharing models that can provide stronger privacy protection than traditional

[8] Wood A, Airoldi E, Altman M, de Montjoye Y-A, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rulemaking. Comments on Regulations.gov. 2016.

de-identification techniques that are in wide use today, including synthetic data, interactive mechanisms, and multiparty computation systems. We further note:

> Many of these data-sharing models are also compatible with a formal privacy guarantee called differential privacy. Differential privacy is a strong, quantitative notion of privacy that is provably resilient to a very large class of potential misuses. As a robust privacy framework that addresses both known and unforeseeable attacks, differential privacy represents a solution that moves beyond the penetrate-and-patch approach that is characteristic of traditional de-identification approaches. We recommend that [government regulations], through the proposed list of approved safeguards, encourage the use of stronger privacy measures, including measures that are compatible with formal privacy models.[9]

Data releases should incorporate more advanced data sharing models, including formal privacy models, where possible, as such techniques can enable wider access and use of data while providing robust privacy protection for individuals.

Thank you for your consideration of these comments.

Respectfully,

Micah Altman
Director of Research, MIT Libraries
Nonresident Senior Fellow, Brookings Institution

[9] *Id.* at 23.