January 6, 2016

Via electronic filing

Jerry Menikoff, M.D., J.D.
Office for Human Research Protections
Department of Health and Human Services
1101 Wootton Parkway, Suite 200
Rockville, MD 20852

Re: Federal Policy for the Protection of Human Subjects; Proposed Rules (HHS–OPHS–2015–0008)

Dear Dr. Menikoff:

We and our colleagues are 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. Through our research, we have developed a number of recommendations that we believe could be incorporated into the Common Rule to enable the wider sharing of research data while providing strong privacy protection for human subjects.

We recognize the exciting research opportunities enabled by new data sources and technologies for collecting, analyzing, and sharing data about individuals. With the ability to collect and analyze massive quantities of data related to human characteristics, behaviors, and interactions, researchers are increasingly able to explore phenomena in finer detail and with greater confidence.[2] At the same time, a major challenge for realizing the full potential of these recent advances will be protecting the privacy of human subjects. Approaches to privacy protection in common use in both research and industry contexts often provide limited real-world privacy protection. We believe institutional review boards (IRBs) and investigators require new guidance to inform their selection and implementation of appropriate measures for privacy protection in human subjects research.

Therefore, we share many of the same concerns and recognize the value of the goals at the heart of the proposed rules to update the Federal Policy for the Protection of Human Subjects (Common Rule):

---

[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] David Lazer et al., *Computational Social Science*, 323 SCIENCE 721 (2009); Gary King, *The Changing Evidence Base of Social Science Research*, *in* THE FUTURE OF POLITICAL SCIENCE: 100 PERSPECTIVES (Gary King, Kay L. Schlozman, & Norman Nie eds., 2009).

> People share information about themselves with large numbers of people with the click of a button, and this trend of rapid and widespread sharing is only likely to grow. The increase in concern about unauthorized and inadvertent information disclosure, in combination with newer research techniques that increase the volume and nature of identifiable data suggest the need for the Common Rule to more explicitly address data security and privacy protection.[3]

The Common Rule is well-positioned to help lead researchers towards state-of-the-art privacy practices as they advance new human subjects research methods and utilize new sources of data. We argue that the Common Rule should address emerging privacy concerns by incorporating definitions informed by recent developments in the scientific understanding of privacy. The Department of Health and Human Services should also provide detailed guidance on choosing among and applying modern data security and privacy techniques. Members of our team have developed a framework for analyzing privacy risks and selecting appropriate controls that are calibrated to the risks and intended uses in specific cases.[4] We believe these concepts could be used to inform the development of a similar framework for the regulation of privacy in human subjects research. Our findings and their applicability to the questions presented in the notice of proposed rulemaking (NPRM) are discussed in detail below.

Brief summary of comments

The NPRM proposes a number of revisions to the Common Rule that are likely to enable increased collection, use, and sharing of personal data for scientific research, while also leading to stronger privacy protection for human subjects. We support the following aspects of the NPRM in particular, as they are likely to both facilitate research and enhance protections for human subjects:

- The development of a regularly-updated list of specific privacy and security measures that would be deemed to satisfy the requirement for reasonable and appropriate safeguards.[5]
- Authorization for IRBs to streamline the case-by-case review of safeguards established by institutions and investigators that have demonstrated compliance with the list of approved safeguards.[6]

[3] Federal Policy for the Protection of Human Subjects; Proposed Rules ("NPRM"), 80 Fed. Reg. 53,933, 53,940 (Sept. 8, 2015).
[4] *See, e.g.*, Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L J. __ (forthcoming), http://privacytools.seas.harvard.edu/files/modernopendataprivacy.pdf; Salil Vadhan et al., Comments to the Department of Health and Human Services Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket No. HHS–OPHS–2011–0005 (Oct. 26, 2011), http://privacytools.seas.harvard.edu/files/commonruleanprm.pdf; Micah Altman et al., Comments to the White House Office of Science and Technology Policy Re: Big Data Study; Request for Information (Mar. 31, 2014), http://privacytools.seas.harvard.edu/files/whitehousebigdataresponse1.pdf.
[5] NPRM at 54,026–27.
[6] *Id.* at 53,988.

- Requirements for implementing safeguards broadly, covering some categories of research activities falling within an exemption to IRB review.[7]

- The use of controls that are calibrated to different categories of data sharing (i.e., in some cases, data shared publicly would be subject to more stringent requirements than data shared among researchers).

At the same time, we identify possible gaps in the scope of coverage of the proposed rules and points where the proposed rules could more closely reflect the latest scholarship on privacy. We find that the proposed rules would exclude some categories of research associated with risks and benefits likely to be similar to those for covered categories of research. In addition, the NPRM refers to a number of common data management concepts and practices, but there is a significant gap between practices in common use and recent advances in privacy. For example, the proposed rules refer to concepts such as de-identification and identifiability, yet traditional de-identification techniques have notable limitations. In addition, definitions based on a binary conceptualization of "identifiability" lack sufficient precision to be used as a general standard. These and related issues suggest that the proposed rules could be revised to direct IRBs and investigators to state-of-the-art practices for privacy protection.

We recommend the development of rules and guidance based on the following principles of a modern approach to privacy:[8]

- Calibrating privacy and security controls to the intended uses and privacy risks associated with the data;
- When conceptualizing informational risks, considering not just re-identification 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 de-identification;
- 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.

In response to the questions presented in the NPRM, we make the following concrete recommendations for incorporating these principles into the regulatory framework for human subjects research protection:

- Use of clear and consistent definitions for privacy, confidentiality, and security, and descriptions of the complementary functions of privacy and security controls.

[7] *Id.* at 53,961–65.
[8] We outline this approach in Micah Altman et al., *Towards a Modern Framework for Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L J. __ (forthcoming).

- Relaxing the sharp binary distinctions between "identifiable" and "non-identifiable" information and between "public" and "private" information.
- Requiring IRBs and investigators to consider the appropriateness of additional privacy and security controls when handling and sharing "de-identified" or "public" information, or information obtained and shared with consent.
- Requiring investigators to conduct systematic evaluations of the risks and intended uses of the data to be generated through a research activity.
- Instructing IRBs and investigators to calibrate the use of privacy and security controls based on a systematic evaluation of the risks and intended uses of the data, and, if appropriate, to implement a tiered access model for closely matching controls to different risks and intended uses at every stage of the information lifecycle.
- Formation of an advisory committee of data privacy experts to help the Secretary of Health and Human Services create and periodically update a list of approved privacy and security safeguards and guidance materials for selecting, applying, and calibrating those safeguards.

General recommendations for updating the Common Rule based on modern privacy principles

The scholarly literature on privacy has revealed significant limitations in many common practices for protecting privacy when collecting, using, and sharing data about individuals. In some key respects, the Common Rule and the proposed revisions in the NPRM rely on traditional conceptions of privacy and techniques for privacy protection, potentially leading to uncertainty, inconsistency, and a suboptimal balancing of privacy and utility. To address these concerns, we have proposed a framework for systematically assessing privacy risks and applying safeguards that are calibrated to the risks and intended uses of the data. As we discuss in the sections that follow, we believe this framework can be used to inform updates to the Common Rule and lead to more consistent, systematic, and appropriate privacy protection for human subjects while enabling increased collection, use, and sharing of data for research.

1. Terminology

Terms such as privacy, confidentiality, security, and sensitivity are used in multiple communities of practice in somewhat different ways, and they are defined inconsistently throughout the literature.[9] This creates uncertainty regarding the contexts in which certain interventions for privacy protection should be applied. Such inconsistency is also found in the NPRM, which refers alternately to "privacy,

[9] For example, the statistical disclosure limitation literature defines "privacy" to refer to the right of data subjects to control the manner and extent of sharing of their information and "confidentiality" to refer to the duty of data holders to prevent unauthorized disclosure after collection. *See, e.g.*, Stephen E. Fienberg, *Confidentiality and Disclosure Limitation*, 1 ENCYCLOPEDIA OF SOCIAL MEASUREMENT 463 (2005). In contrast, the literature on cryptography often uses "privacy" to refer to controls over disclosure or to the absence of a privacy breach, *see, e.g.*, Cynthia Dwork, *Differential Privacy*, ENCYCLOPEDIA OF CRYPTOGRAPHY AND SECURITY 338 (2011), and the information security literature uses the term "confidentiality" to refer to controls over disclosure but in the narrower context of an information system, *see, e.g.*, RICK LEHTINEN, DEBORAH RUSSELL, & G.T. GANGEMI, SR., COMPUTER SECURITY BASICS 197 (2006).

confidentiality, and security protections," "data protection standards," "information security measures," "data security and information privacy protections policies," and "data security and information protection standards," among other related terms.[10] It is not clear whether the use of different terms is intended to reflect meaningful distinctions or whether some of the terms are used interchangeably.

The Common Rule and related policy should be revised to include clear definitions, descriptions of appropriate privacy and security measures, and detailed implementation guidance for IRBs and investigators. It should provide definitions for terms such as privacy, confidentiality, security, and sensitivity, and these terms should be used consistently throughout the regulations. These definitions should be developed based on a modern understanding of privacy that is not limited to a binary conception of "identifiability" but covers more broadly the potential for others to learn about individuals based on the inclusion of their information in the data. For example, as discussed in more detail below in Section 6, a privacy standard could require "the protection of individuals from the possibility that their personal information would be directly revealed or otherwise inferred." Such a definition would take into account not just whether an individual can be directly associated with a particular attribute, but more broadly the extent to which attributes that may be revealed or inferred depend on an individual's data and the potential harm that may result. It would also provide a goal against which privacy measures, including emerging formal privacy models, could be evaluated.

We suggest definitions for these terms as they are used in these comments. Our intent is not to privilege one field's use of these terms over another generally, but to provide consistency and clarity. Briefly, we use "information security" roughly as it is defined in the information security field and in the Federal Information Security Management Act[11] to mean the protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability. We recommend resources from NIST as references for information security related terminology, and for typologies of information security controls, and we recommend that "information security" be used consistently to refer to protection of information within organizations and information systems.[12] Similarly, in the same spirit as references from NIST, we use "privacy" to refer to a set of concerns interrelated with security but separate from it, and which is broadly inclusive of individuals' awareness of, consent to, control over, and exposure to risks from the collection, storage, dissemination, and use of information generated from observation or interaction with them.[13] Finally, we use "sensitivity" as we have in prior work to refer to a summary of the extent, type, and

[10] *See* NPRM at 53,954 ("privacy, confidentiality, and security protections"), 53,978 ("privacy and confidentiality protections"), 53,968 ("data protection standards"), 53,980 ("information security measures"), 53,940 ("data security and privacy protection"), 53,937 ("data security and information privacy protections policies"), 53,955 ("data security and information protection standards").

[11] Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–49 (2013).

[12] *See, e.g.*, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS, Special Publication 800-53 (4th rev. 2013), http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf.

[13] *See id.*

likelihood of harms that could result when a privacy threat is realized;[14] thus, sensitivity is inherently derived from harms and threats, as defined in Section 3 below.

Distinctions between terms such as privacy and security are particularly relevant for section 105 of the proposed rules, which would require investigators to "implement and maintain reasonable and appropriate safeguards."[15] Section 105 should acknowledge the complementary roles of privacy and security controls, where security controls can be viewed as restricting access to information and privacy controls as limiting the potential for harm once access to information is granted.[16] This section should also require IRBs and investigators to consider the suitability of both security and privacy controls in their data management programs,[17] and specify that there is a wide range of procedural, economic, legal, educational, and technical controls for protecting security and privacy.[18] It should clarify that no single solution is appropriate in all cases; rather, the selection of privacy and security controls should be calibrated to the specific risks and intended uses of the data. As discussed below in Sections 3 through 5, the Secretary of Health and Human Services, in consultation with an advisory committee of data privacy experts, should develop detailed guidance materials to help IRBs and investigators systematically evaluate the privacy risks associated with their activities, as well as choose privacy and security controls that are suitable for mitigating those risks while enabling the data uses they intend to support. This guidance should also emphasize the complementary roles of security and privacy controls, advise IRBs and investigators that security controls are necessary but not sufficient conditions for protecting the privacy of human subjects, and advise investigators on implementing a combination of appropriate privacy and security controls.

2. Recognition of the limitations of de-identification

Advances in the scientific understanding of privacy have demonstrated that privacy measures in common use, such as de-identification, have significant limitations. We argue that de-identification should no longer be used as a general standard for privacy protection in the absence of a systematic analysis of

---

[14] Micah Altman et al., *Towards a Modern Framework for Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L J. __ (forthcoming).

[15] NPRM at 54,049–50.

[16] For further discussion of the difficulty of defining and achieving privacy even given perfect cryptographic and access control mechanisms, see Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMUNICATIONS OF THE ACM 86 (2011).

[17] Resources developed by the National Institute of Standards and Technology (NIST) could serve as a model for applying a distinction in implementation guidance on appropriate privacy and security controls. For example, NIST defines security controls as encompassing safeguards within information systems and their environments to protect information during processing, storage, and transmission. Security controls fall within categories such as access, identification and authentication, and system and information integrity controls, among others. In contrast, privacy controls are defined as administrative, technical, and physical safeguards to protect and ensure the proper handling of information associated with privacy risks. Categories of privacy controls include authority and purpose, data minimization and retention, individual participation and redress, transparency, and use limitation controls. *See* NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS, Special Publication 800-53 (4th rev. 2013).

[18] See Table 1 below for a sample of the types of controls that should be considered for inclusion in the list of safeguards to be developed in accordance with section 105 of the proposed rules.

informational risks and intended uses, and, in many cases, the implementation of additional privacy and security controls. We highlight the following concerns in particular:

1. Approaches to privacy that rely only on protection through de-identification, while they may reduce some risks, often do not prevent all disclosures, do not minimize risks to individuals, or do not protect information in the manner that most individual subjects would expect.[19] De-identified data can, in many cases, be re-identified easily. For instance, numerous re-identification attacks have demonstrated that it is often possible to identify individuals in data that have been stripped of direct and indirect identifiers.[20] It has been shown more generally that very few pieces of information can be used to uniquely identify an individual in a released set of data.[21]

2. De-identification generally ignores many types of attacks and attackers. Approaches to de-identification in common use are typically designed to address only a subset of disclosure risks. For instance, common techniques focus on removing or aggregating certain categories of information, such as addresses and dates of birth, which can often be easily linked to information from other sources. These approaches do not address risks that are associated with attributes not specified as sensitive by the individual applying the technique, risks that are not based on the direct linkage of attributes across different sets of data, or risks that an employer, insurance company, relative, or friend may have extensive knowledge about the subject that could be used for re-identification. Furthermore, the traditional notion of identification as a binary determination overlooks attribute inference, or the ability to infer a characteristic of an individual with some probability based on the inclusion of that individual's information in the data.[22] Where traditional de-identification techniques provide estimates of re-identification risk, these estimates should be interpreted as weak lower bounds that depend strongly on assumptions made about what the attacker might know (e.g., whether the attacker knows if the target individual is in the dataset or not), rather than rigorous measures of risk.

3. De-identification is not readily scalable. De-identification techniques, if applied by expert statisticians, can sometimes provide reliable privacy protection. For example, statistical agencies are equipped to apply sophisticated disclosure limitation techniques to mitigate privacy risks before releasing data to the public. However, the techniques they use cannot readily be applied effectively by non-experts, and it is not reasonable to require expert statisticians to be involved in every data release covered by the Common Rule.

---

[19] *See* Arvind Narayanan & Edward W. Felten, *No Silver Bullet: De-identification Still Doesn't Work* (2014), http://www.randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

[20] *See* Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J. L., MED., & ETHICS 98; Latanya Sweeney, *Uniqueness of Simple Demographics in the US Population*, Data Privacy Lab Technical Report (2000).

[21] *See* Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 Science 536 (2015); Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 Nature Sci. Rep. 1376 (2013).

[22] *See, e.g.*, Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, & Alex (Sandy) Pentland, *Predicting Personality Using Novel Mobile Phone-based Metrics*, PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING, BEHAVIORAL-CULTURAL MODELING, AND PREDICTION 48 (2013).

4. De-identification often results in the redaction or withholding of useful information. Traditional de-identification techniques often reduce the quality of data that are released, limit the scope of research questions that can be answered using the data, or yield results that support spurious correlations or erroneously indicate a lack of statistical significance. For example, the HIPAA Privacy Rule safe harbor method for de-identification requires the removal of all location information (among other pieces of information), except for the state or the first three digits of a ZIP code for geographic areas with a population greater than 20,000 people. Data released using this approach cannot be used to investigate trends at the county level. However, in some cases, the information removed using heuristic de-identification methods could be safely shared using alternative methods. For instance, if the data in this example were released through contingency tables instead of redacted individual-level data, the data could potentially be used to study trends at the county level while providing strong privacy protection, especially if the contingency tables were designed to satisfy a formal guarantee of privacy such as differential privacy.[23]

5. Framing regulation around de-identification and drawing sharp distinctions between "identifiable" and "non-identifiable" information encourage the use of de-identification techniques and hinder adoption of stronger privacy measures. For instance, the safe harbor method for de-identification described by the HIPAA Privacy Rule equates privacy with simple redaction of direct and indirect identifiers and endorses this technique as adequate privacy protection. However, heuristic de-identification approaches should not be considered generally sufficient in the absence of a systematic analysis of risks and the consideration of the suitability of additional privacy and security controls. Moreover, a regulation that relies on a binary conceptualization of identifiability provides little guidance for the use of emerging formal privacy-preserving approaches for which the terminology does not apply.

We recommend that the Common Rule be updated to recognize the limitations of de-identification. It should not refer to "identifiability" as a strict binary determination and broadly exclude "non-identifiable" information, as these terms lack precision and their use encourages investigators to employ heuristic de-identification techniques such as redaction of pieces of information deemed to be identifying. We also argue that the Common Rule should not endorse the exclusive use of de-identification techniques, such as the HIPAA Privacy Rule safe harbor method for de-identification, without requiring investigators to conduct a systematic analysis of privacy risks and implement additional privacy controls as appropriate. IRBs and investigators should receive detailed guidance on the limitations of common de-identification methods and be encouraged to implement additional privacy and security controls to mitigate the risks not addressed by de-identification.

The Common Rule should also require that human subjects be informed of the privacy and security controls that will be used to protect their data. Consent forms should clearly describe the limitations of the measures used and the risks that may remain despite the safeguards put in place. For instance, the proposed language in section 116 requiring consent forms to disclose whether the data contain

[23] For a discussion of alternatives to traditional de-identification techniques, see the discussion of advanced data-sharing models and emerging formal privacy methods in Section 8 below.

"identifiable information"[24] should be revised to require disclosure of the specific steps that have been taken to mitigate privacy risks as well as the risks to individuals that may remain despite these efforts.

3. The need for a systematic evaluation of the context of intended uses and privacy risks

We recommend that IRBs and investigators be required to conduct a systematic analysis of the intended uses and privacy risks associated with a research activity. In particular, this requirement could be incorporated in section 105 of the proposed rules, which requires investigators to implement privacy and security controls that are reasonable and appropriate for a specific research activity. We recommend that IRBs and researchers be required to comprehensively assess factors related to the context, uses, threats, harms, and vulnerabilities that may be associated with a particular research activity.[25] Definitions for these concepts and how they can be applied in a privacy analysis are discussed in turn below.

Context and purpose. We argue generally that the selection of appropriate safeguards should take into account the purpose behind the research activity, the context in which the data were collected, and expectations and norms regarding the use of the data. This is an approach reflected in broadly-applied frameworks such as the fair information practices[26] and contextual integrity,[27] and related to the ethical principle of respect for persons set forth in the Belmont Report.[28] We recommend that these considerations inform revisions to the Common Rule generally and form a component of a required systematic assessment of risks and intended uses. For instance, the NPRM proposes exclusions for research using information that has been or will be acquired solely for non-research activities, as well as information that was acquired for research studies other than the proposed research study when the sources are publicly available.[29] For these categories of research, it is likely that an investigator's use of the information will differ from the subjects' expectations with respect to their information. Investigators should be required to evaluate the context in which this information was collected as well as the subjects' expectations regarding its use, when determining whether certain privacy and security controls, such as use limitations, could be implemented to protect the subjects. The Secretary of Health and Human Services should develop guidance to help IRBs and investigators assess the expectations of and uses intended by the subjects, as well as the threats, harms, and vulnerabilities associated with the potential research uses.

Utility and intended uses. Investigators should be required to take into account the intended primary and secondary uses of the research data. Investigators sharing data and those who might seek to use it in the

---

[24] *See* NPRM at 54,018.
[25] *See* Micah Altman et al., *Towards a Modern Framework for Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L J. __ (forthcoming).
[26] *See* ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, THE OECD PRIVACY FRAMEWORK (2013), http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.
[27] *See* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119 (2004).
[28] NATIONAL COMMISSION FOR THE PROTECTION OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RESEARCH, BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH (1978).
[29] *See* NPRM at 53,952–53.

future often have certain uses in mind, such as conducting individual-level vs. population-level analyses, or linking the released information with other data sources. A data management decision affects the output of the data, such as whether the data are made available through raw individual-level data, a summary table, an interactive query interface, model parameters, or a static or dynamic visualization, among other alternatives. Similarly, the type of methodology desired by a researcher can vary between contingency tables, summary statistics, regression models, data mining, and other analysis types. These choices affect the privacy controls that may be suitable, such as differential privacy, de-identification, and secure data enclaves, among others. Evaluating the utility of the data to be collected, used, or released involves an evaluation of the types of uses or analytic purposes intended by each of the relevant stakeholder groups, and how privacy controls implemented at each stage enable or restrict such uses. Because these are factors that will vary in different research activities, requiring one technique for privacy protection, such as de-identification, is not appropriate for all cases. Rather, the selection of controls in a given case should be calibrated to the specific uses, threats, harms, and vulnerabilities associated with it.

Threats. Investigators should also be required to systematically assess the privacy threats associated with a research activity. Threats are defined broadly as potential adverse circumstances or events that could cause harm to a data subject as a result of the inclusion of that subject's data in a specific data collection, storage, use, or release. Threats include everything from government surveillance, to a researcher losing a laptop with the data, to natural disasters. The concept encompasses factors related to the capabilities of adversaries and the sensitivity of the information. Modeling adversaries, either formally or informally, typically involves specifying their objectives, the auxiliary knowledge they possess, and their resources or capabilities.[30] Some broad examples of potential adversaries include nosy neighbors or relatives, former spouses, business competitors, data brokers, muckraking journalists, potential employers or insurers, oppressive governments, and others. In each of these examples, the adversary will have distinct goals (such as targeting a specific individual in the data or any individual in the data), knowledge (such as extensive personal knowledge or demographic information gleaned from third-party sources), and capabilities (consider the resources of a large data broker relative to that of a nosy neighbor).[31]

Harms. Investigators should also evaluate privacy harms, or injuries sustained by data subjects as a result of the realization of a privacy threat. There is a broad range of informational harms recognized by regulation and by researchers in the behavioral, medical, and social science fields. Examples include loss of insurability, loss of employability, market discrimination, criminal liability, psychological harm, loss of reputation, emotional harm, and loss of dignity. Harms to groups and society include social harms to a vulnerable group such as stereotyping, price discrimination against vulnerable groups, market failures (e.g., by enabling manipulation, or eliminating uncertainties on which insurance markets are predicated), as well as broad social harms such as the chilling of speech and action, potential for political discrimination, or blackmail and other abuses. Investigators should consider the sensitivity of the data, or characteristics of the data related to the extent, type, and likelihood of harms that could result when a threat is realized. Generally, information should be treated as sensitive when that information, if linked to

[30] For a general detailed and thoughtful discussion of threat models in the privacy context, see Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117 (2013).

[31] *See* LEON WILLENBORG & TON DE WAAL, ELEMENTS OF STATISTICAL DISCLOSURE CONTROL (2001).

a person, even partially or probabilistically, possibly in conjunction with other information, is likely to cause significant harm to an individual, group, or society. Harms may occur directly as the result of a reaction by a data subject or third party to the information, or indirectly as a result of inferences made from information. An example of a potential harm that is indirect and inferential but nevertheless substantial is the recent demonstration that Facebook "likes" can be used to "automatically and accurately predict a range of highly sensitive personal attributes including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender."[32] A set of data may, therefore, be very sensitive and have the potential to cause serious harm, even if it does not contain pieces of information that have traditionally been considered sensitive.

Vulnerabilities. Privacy vulnerabilities are defined as characteristics that increase the likelihood that threats will be realized. These characteristics are defined as broadly inclusive, encompassing characteristics of the data; of the systems used to collect, store, manage or release the data; and of the related context in which these systems operate and in which interactions with these systems occur. They may arise from the characteristics of the data being collected, managed, or released; of the logical or physical systems used to manage that data; or of the broader context of release. Vulnerabilities are associated with the scope of information collected, maintained, used, and disseminated by an investigator. They also encompass the "identifiability" of the data or the potential for learning or inferring attributes about individuals based on the inclusion of those individuals' information in the data. It is important to note that vulnerabilities may remain after the suppression of information deemed to be directly or indirectly identifying. For example, vulnerabilities may support indirect linkages to auxiliary information, statistical re-identification, learning about individuals without identifying them (e.g., "attribute disclosure"), or learning about characteristics of specific groups. For this reason, privacy and security controls should often be used in combination with traditional de-identification techniques.

Investigators should be required to systematically assess the above factors related to context and purpose, utility and intended uses, threats, harms, and vulnerabilities. The privacy and security controls that should be considered reasonable and appropriate for a specific research activity will vary depending on these characteristics. In the following section we present a conceptual framework for aligning privacy and security controls with the context, intended uses, threats, harms, and vulnerabilities in a specific case. We recommend that a framework similar to the one outlined below be developed by the Secretary of Health and Human Services in consultation with an advisory committee of data privacy experts and incorporated into study design and IRB review processes.

4. Calibration of privacy controls to the specific risks and intended uses

The NPRM seeks "to create information privacy protections that would apply to research, calibrated to the level of identifiability and sensitivity of the information being collected."[33] As outlined in the previous section, we agree that such calibration should be a requirement for implementing privacy and security

[32] *See* Michal Kosinski et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 5802 (2013).
[33] NPRM at 53,978.

safeguards that are reasonable and appropriate. In this section, we present a suggested framework for tailoring privacy and security controls to the context, intended uses, threats, harms, and vulnerabilities associated with a specific research activity.[34] We recommend that a similar framework be developed by the Secretary of Health and Human Services, in consultation with an advisory board of data privacy experts, to guide investigators through a systematic analysis and selection of appropriate privacy and security controls.

Advances in science and technology enable the increasingly sophisticated characterization of privacy risks and harms and new interventions for protecting data subjects. For applying these concepts and tools in practice, we propose 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/release, and post-access stages. As outlined in the previous section, we recommend that investigators be required to identify the context and expectations of data subjects, specify the desired data uses and expected benefits, and examine each stage of the data lifecycle to identify specific privacy threats, harms, and vulnerabilities. The Secretary of Health and Human Services should provide guidance to IRBs and investigators for selecting among a distinct set of legal, technical, economic, procedural and educational interventions at each stage of the information lifecycle based on the specific context, intended uses, threats, harms, and vulnerabilities associated with the research activity. We also recommend that the list of privacy and security controls to be developed under section 105 of the proposed rules take a similar approach to cataloging the broad range of interventions available and not focus on a narrow set of security controls such as encryption and access control. We provide an example of such a catalog below in Table 1 in Section 8. Implementation guidance should also describe the contexts in which each control should be considered reasonable and appropriate.

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. 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.[35] The y-axis provides a scale for the post-transformation identifiability, the potential for others to learn about individuals based on the inclusion of their information in the data, and 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.

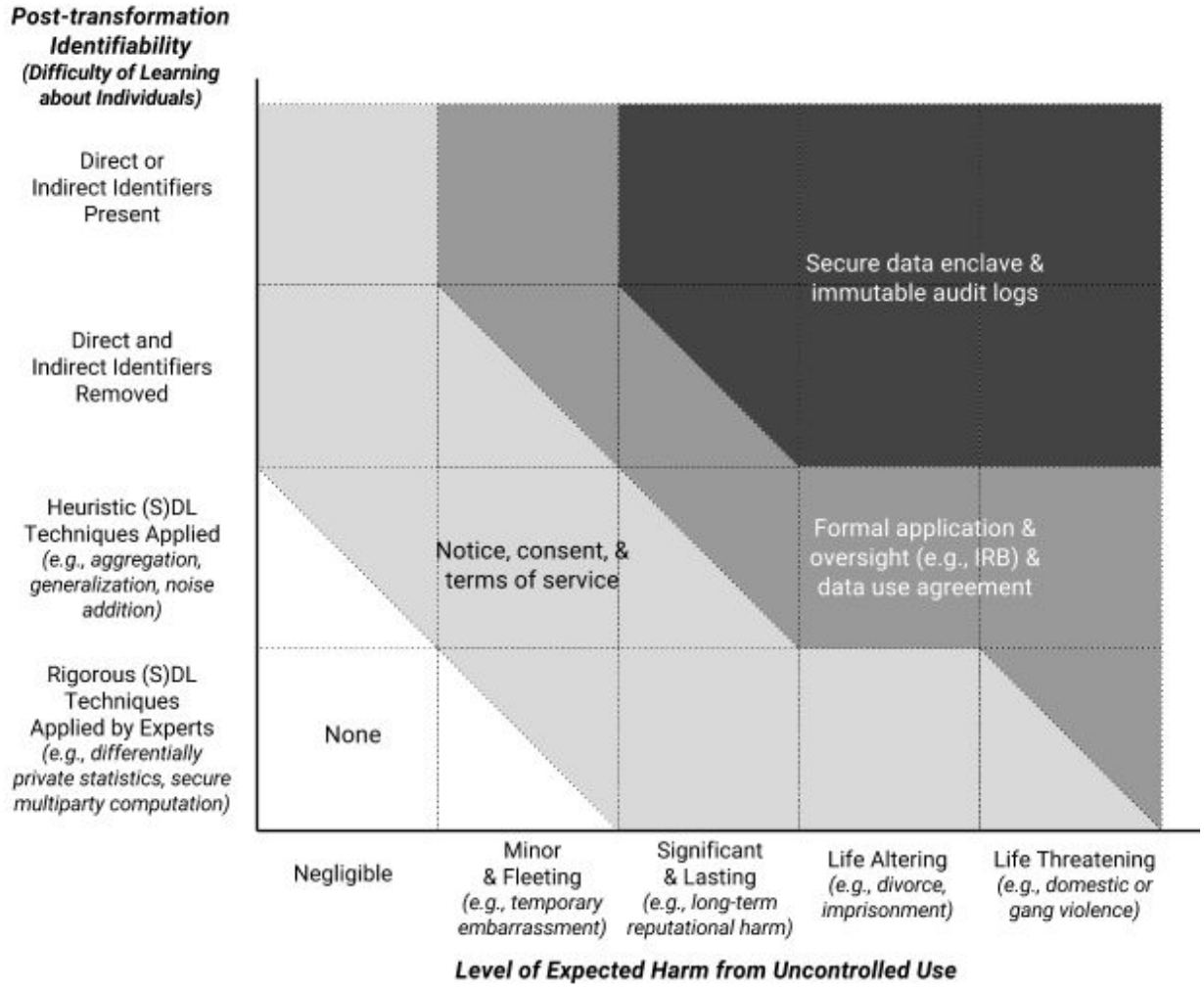
---

[34] *See* Micah Altman et al., *Towards a Modern Framework for Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L J. __ (forthcoming).

[35] Although their guidance on measuring harm is incomplete, there is a rough consensus in the IRB literature on typologies of harm and an approximate ranking of many common risks. *See, e.g.*, ELIZABETH A. BANKERT & ROBERT J. ANDUR, INSTITUTIONAL REVIEW BOARD: MANAGEMENT AND FUNCTION (2006); RAYMOND M. LEE, DOING RESEARCH ON SENSITIVE TOPICS (1993).

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 Section. 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.

Figure 1. Calibrating privacy and security controls.



For many research 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 IRBs

and investigators be encouraged to 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 IRB, 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 IRBs and investigators calibrate the 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.

5. Formation of an advisory committee of data privacy experts

The NPRM notes that "IRBs were not designed to evaluate risks to privacy and confidentiality, and often have little expertise in these matters. Setting uniform specific standards will help to assure appropriate privacy and confidentiality protections to all subjects, without the administrative burden of needing a specific committee review of the privacy and confidentiality protections of each study."[36] We agree that the full burden of evaluating privacy risks and selecting appropriate safeguards should not rest on IRBs and that the Common Rule should promote consistency in treatment of privacy risks across different research studies. However, there is no uniform set of safeguards that is appropriate for all settings. As discussed above in Sections 3 and 4, privacy and security controls should be calibrated to the specific context, intended uses, threats, harms, and vulnerabilities associated with a given research activity. Given the complexity of a privacy analysis and the rapidly changing state of the art for privacy-preserving techniques, we recommend that the Common Rule establish an advisory committee of data privacy experts and other stakeholders to develop detailed guidance for IRBs and investigators.

We recommend that members of the advisory committee include data privacy experts from computer science, statistics, and law; IRB administrators; regulators; and human-subjects researchers. An agency with technical expertise such as the National Institute of Standards and Technology should oversee this process.[37] Given the rapidly evolving nature of advances in the field of privacy, the advisory committee should convene regularly to update the guidance, every two to five years (with greater frequency in the near term).

The Secretary of Health and Human Services should consult with the advisory committee to develop and regularly update guidance materials covering the following topics:

---

[36] NPRM at 53,978.

[37] In fact, NIST is currently engaged in a comprehensive review of de-identification and other privacy measures, and the advisory committee we propose could benefit from these related efforts. *See* SIMSON L. GARFINKEL, DE-IDENTIFICATION OF PERSONAL INFORMATION, National Institute of Standards and Technology Internal Report 8053 (Oct. 2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf; NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, PRIVACY RISK MANAGEMENT FOR FEDERAL INFORMATION SYSTEMS, Internal Report 8062 (Draft) (May 2005), http://csrc.nist.gov/publications/drafts/nistir-8062/nistir_8062_draft.pdf.

- Definitions and instructions for interpreting relevant terminology such as privacy, confidentiality, and security, and a guide to the complementary roles of privacy and security controls.

- Factors to consider in a systematic privacy analysis, including criteria related to the characteristics of information, systems, actors, and activities relevant to the context, intended uses, threats, harms, and vulnerabilities at each stage of the information lifecycle.

- A catalog of privacy and security controls, including a wide range of procedural, legal, technical, economic, and educational controls, for protecting privacy and security at each stage of the information lifecycle.

- Annotations for each privacy or security control that describe specific types of information, uses, and contexts in which the given control should or should not be considered appropriate.

- Criteria for selecting combinations of privacy and security controls that are appropriate based on the risks and intended uses in a specific case.[38]

- Guidelines for disclosing in consent forms which privacy and security controls have been or will be applied, and the vulnerabilities they do and do not address.

To incorporate this approach into the Common Rule, we suggest the following changes to the proposed rules. We recommend that language establishing an advisory committee of data privacy experts be added in the final rule. Section 105 should be revised to provide that the list of approved safeguards be developed not only by the Secretary of Health and Human Services in consultation with other federal agencies and departments, but also with this advisory committee of data privacy experts.

We also recommend that section 105 clarify that the list of measures deemed to satisfy requirements for reasonable and appropriate safeguards should not serve as a checklist of measures to be implemented by all investigators in all cases. Rather, the list and accompanying implementation guidance should emphasize that the safeguards that are appropriate for a given research activity should be calibrated to the specific risks and intended uses associated with a particular case. The implementation guidance should include guidance on such calibration in order to promote consistent application and reduce the need for case-by-case evaluation by IRBs. The final rule should require the Secretary, in consultation with the advisory committee, to develop guidance materials for implementing these safeguards and calibrating them in specific contexts, using a framework similar to the one we propose in Sections 3 and 4 above. In addition, we recommend that use of controls such as the HIPAA Privacy Rule safe harbor method for de-identification not be written into the Common Rule as an exemption, but that such controls be considered for inclusion in the list of safeguards that will be regularly reviewed and updated over time.

Responses to enumerated questions in the NPRM

[38] For an example that illustrates how guidance on calibrating appropriate controls to privacy risks and intended uses can be constructed, see Figure 1 and the related discussion below.

6. Definition of identifiable private information

Question 3. To what extent do the issues raised in this discussion suggest the need to be clearer and more direct about the definition of identifiable private information?

Response: The NPRM proposes to retain the Common Rule's current standard of identifiability without modification.[39] It would also continue to permit, without consent, the secondary research use of non-identified private information, such as medical records that have been redacted in accordance with the HIPAA Privacy Rule safe harbor method for de-identification.[40] As discussed above in Section 2, identifiability should not be considered a strong binary determination and therefore the scope of privacy protection should not depend on a classification of information as "identifiable" or "non-identifiable." In addition, the Common Rule's emphasis on concepts such as identifiability and "identifiable" information encourages researchers to limit their choice of privacy controls to heuristic de-identification techniques, despite the availability of many other controls that are potentially more effective.

For these reasons, we recommend that the binary identifiability standard be replaced with a clearer statement of a privacy goal. Instead of excluding "non-identifiable" information, the regulations should require the implementation of privacy and security safeguards that are proportional to the expected harm from learning about an individual in the data. We propose the following as possible alternative standards:

- "Reasonable and appropriate safeguards should be implemented to protect individual subjects from the possibility that their personal information would be directly revealed or otherwise inferred."[41]

- "No individual should incur more than a minimal risk of harm from the use of his or her data in computing the values to be released, even when those values are combined with other data that may be reasonably available."[42]

In other words, the standard should not depend on a notion of privacy risk that considers only whether an individual can be directly associated with a particular attribute. Rather, the standard should take into account the extent to which attributes that may be revealed or inferred depend on an individual's data and

[39] NPRM at 53,945.
[40] NPRM at 53,974.
[41] By referring to "personal" information, we mean to exclude information about individuals that can be deduced from combining scientific knowledge about the population as a whole and prior knowledge about that individual, even if that information might be considered sensitive. Indeed, the point of research is to learn about populations, and such knowledge necessarily will also teach us about individuals. Instead, here we refer only to information that would not have been revealed had the specific individual been omitted from the study (in the spirit of differential privacy). *See* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 1 COMMUNICATIONS OF THE ACM 86 (2011). This approach is consistent with one of the principles established by the Common Rule: an IRB "should not consider possible long-range effects of applying knowledge gained in the research." 45 C.F.R. § 46.111.
[42] Salil Vadhan et al., Comments to the Department of Health and Human Services Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket No. HHS–OPHS–2011–0005 (Oct. 26, 2011), http://privacytools.seas.harvard.edu/files/commonruleanprm.pdf.

the potential harm that may result. The language we propose also provides a goal against which privacy measures, including emerging formal privacy models, can be evaluated. This will be critical to encouraging the use of robust measures for privacy protection such as privacy-aware methods for producing contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations, discussed below in Section 8.

Alternatively, should the Common Rule continue to draw binary distinctions between "identifiable" and "non-identifiable" information, it should at minimum require investigators and research subjects to be informed of the privacy risks that may be associated with data deemed to be "non-identifiable." Subjects should be informed of the privacy and security controls in place and the risks that may remain. Investigators should also be provided with guidance for taking these risks into account and, if appropriate, implementing additional privacy and security controls when handling "non-identifiable" data. In particular, should the Common Rule continue to permit the disclosure of data de-identified according to the HIPAA Privacy Rule safe harbor method without consent, it should direct IRBs and investigators to consider the suitability of additional privacy and security controls to mitigate risks not addressed by redacting certain identifiers. In addition, we recommend that de-identification methods such as the safe harbor method not be specifically referenced in the regulation as sufficient safeguards for complying with the Common Rule. Instead, we recommend that de-identification methods be included separately in the list of approved privacy and security safeguards to be reviewed and regularly updated in accordance with section 105.

7. Scope of coverage and consistency in application of privacy protections

Research oversight should aim to cover the scope of human subjects research and ensure that similar privacy risks are treated similarly. There are gaps in the Common Rule where different categories of research are subject to different rules although the risks to human subjects may be similar. For instance, there are potential gaps in coverage for research involving many categories of information currently deemed to be public or non-identifiable; privately funded research; and research activities across all stages of the lifecycle, including storage, processing, analysis, release, and post-release. Although these categories may be treated differently for legal, political, or economic reasons, analysis of privacy risks in these area should be based on the same scientific principles discussed above. We recommend that the Secretary of Health and Human Services, in consultation with the advisory committee of data privacy experts proposed above in Section 5, seek to improve consistency in treatment of privacy risks. In particular, guidance should be developed to help IRBs and investigators calibrate privacy and security controls to the context, intended uses, threats, harms, and vulnerabilities in specific cases.

Question 16. Public comment is sought regarding whether it is reasonable to rely on investigators to make self-determinations for the types of research activities covered in this particular exclusion category. If so, should documentation of any kind be generated and retained?

Question 17. Public comment is requested on the extent to which covering any of these activities under the Common Rule would substantially add to the protections provided to human research subjects. Is

there a way in which this exclusion should be narrowed? Public comment is also sought regarding whether activities described here should appear as an exclusion or as an exemption.

Questions 16 and 17 refer to the NPRM's proposal to exclude from coverage "research involving the collection or study of information that has been or will be acquired solely for non-research activities or was acquired for research studies other than the proposed research study when the sources are publicly available, or the information is recorded by the investigator in such a manner that human subjects cannot be identified, directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects or otherwise conduct an analysis that could lead to creating individually identifiable private information."[43] The NPRM relies on the following rationale for this exclusion: "(1) the information is already available to the public, and so any risk it may include exists already, or (2) the information recorded by the investigator cannot be identified, and no connection to or involvement of the subjects is contemplated."[44]

Response: This approach potentially excludes some research activities that pose risks similar to those associated with covered research activities. We caution that risk of harm to human subjects increases with each use of their information, even if the information is available through other sources. Further, the distinction between "public" and "private" information is not a strong binary determination and is the subject of significant debate.[45] For instance, types of information collected from postings on social media, or by sensors in public spaces, may be deemed to be publicly available information, and excluded from the Common Rule, even though they contain sensitive details about individual human behavior. The Secretary's Advisory Committee on Human Research Protections has developed draft guidance on the use of data collected from Internet sources, which takes some first steps at addressing the difficult issues associated with determining whether information collected online qualifies as public or private under the existing regulations.[46] However, what is considered to fall within this definition is open to interpretation and will likely evolve over time. In addition, as discussed above in Sections 2 and 6, demonstrating that information "cannot be identified" is quite difficult. This language could be interpreted to endorse, as a sufficient practice, heuristic de-identification techniques that often do not provide strong privacy protection.

For these reasons, we recommend that the Common Rule not exclude all "public" information or information that "cannot be identified." Instead, we recommend considering an exemption for research activities for which privacy and security safeguards have been implemented in satisfaction of one of the standards we propose in Section 6. We recommend that the framework we outline above in Sections 3 and 4 be incorporated into guidance on implementing the privacy safeguards from section 105 of the proposed rules and serve as a guide for selecting safeguards for research activities subject to this exemption. We

[43] NPRM at 53,952–53.
[44] NPRM at 53,953.
[45] For a survey of approaches to distinguishing between public and private information, see David R. O'Brien et al., *Integrating Approaches to Privacy Across the Research Lifecycle: When Is Information Purely Public?*, Berkman Center Research Publication 2015-7 (2015), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2586158.
[46] Secretary's Advisory Committee on Human Research Protections (SACHRP), Considerations and Recommendations of Concerning Internet Research and Human Subjects Research Regulations, with Revisions (March 2013).

believe it is reasonable to rely on investigators to make self-determinations regarding the standards we propose in Section 6. However, we recommend that the Secretary develop detailed guidance on interpreting these standards. Requiring documentation of the rationale supporting a self-determination is also a reasonable approach.

Question 49. Public comment is sought on the types of research that should fall under the proposed exemption. Should the proposed exemption be available to all types of research using identifiable data collected for non-research purposes or should the exemption be available only to a more limited subset of research? For example, should the proposed exemption apply only for research using records and information already subject to comprehensive privacy and other protections in other Federal laws (e.g., records held by the Federal Government subject to the Federal Privacy Act, or records governed by HIPAA or FERPA)?

Response: There are categories of research using identifiable data collected for non-research purposes that would likely carry risks to human subjects similar to those associated with research using data originally collected for research purposes. Further, the proposed exemption appears to assume that investigators will be able to calibrate the privacy controls they choose to the specific risks and intended uses of the data, without guidance or review from an IRB. In practice, we believe IRBs and researchers would require guidance on selecting among and implementing privacy measures that are appropriate to the specific contexts in which they collect, store, use, and share data.

We thus recommend that the framework we outline above in Sections 3 and 4 be incorporated into guidance on implementing the privacy safeguards from section 105 and serve as a guide for selecting safeguards for research activities subject to this exemption. We recommend that IRBs and investigators be provided with detailed guidance materials that aim to achieve consistency in the treatment of similar privacy risks associated with research activities, whether the information was originally collected for research purposes or otherwise. Such guidance should be developed to help IRBs and investigators evaluate the threats, harms, and vulnerabilities associated with a specific research activity based on factors related to the sensitivity and identifiability (or learning potential from use) of the information.

We note that the proposed exemption would apply as long as prior notice had been given to subjects that their information may be used in research, the privacy and security safeguards of section 105 are required, and the information is used only for the specific research purposes for which the investigator requested access. We recommend that, even under these conditions, IRBs should be empowered to require specific selection of privacy controls based on the guidance developed in consultation by the advisory committee and its evaluation of risk in a given research proposal. In addition, IRBs should be permitted to forgo individual review once establishing such requirements for a class of research activities associated with similar risks and intended uses. Also, while we agree that requiring controls such as notice is a reasonable approach, we recommend that an IRB evaluate and review the sufficiency of the notice and privacy safeguards provided. For instance, we argue that adequate notice should include at least implied consent, and it should not be applied retroactively. We therefore recommend that the Secretary develop specific guidance for providing adequate notice. The Secretary should also consult with the advisory committee of

data privacy experts proposed in Section 5 to develop guidance for the selection of additional information accountability and other controls based on privacy risks from data retention and dissemination.

8. Developing a list of reasonable and appropriate privacy and security controls

Question 71. Public comment is sought regarding whether particular information security measures should be required for certain types of information or research activities and, if so, what measures and for what types of information or research. Specifically, should the safeguards be calibrated to the sensitivity of the information to be collected?

Response: Yes. We agree that particular information security measures should be required for certain types of information or research activities. We recommend that the list of approved safeguards be developed and regularly updated in consultation with the advisory board of data privacy experts we propose above in Section 5. As discussed in Sections 3 and 4, we also recommend that section 105 of the proposed rules require a systematic analysis of the risks and intended uses of the data and consideration of the suitability of a wide range of privacy and security controls. The rules should be accompanied by detailed guidance on selecting and calibrating the controls based on the risks and intended uses in a particular context.

Hence, we agree with the approach taken in section 105 of the proposed rules to provide a list of approved privacy and security measures. As noted above in Section 7, we also support the proposed requirements for implementing reasonable and appropriate safeguards for a wide range of research activities, including certain categories of research subject to an exemption to IRB review. We recommend that this list include not just a narrow set of information security controls but a more comprehensive range of the privacy and security controls that are available across the entire information lifecycle, from the collection stage through post-release stages. 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 and should be considered for inclusion in this list.

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;	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

	Purpose specification; Privacy impact assessments				
Transformation	Process for correction		Metadata; Transparency	Right to correct or amend; Safe harbor de-identification 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

The list should be regularly updated, though we recommend that the review occur more frequently than every eight years as provided in the proposed rules. We recommend that this list be reviewed every two to five years (with greater frequency in the near term). Such an approach would accommodate types of research and data sharing not previously anticipated, as well as new developments in our understanding of threats and approaches to data privacy. In addition, we recommend that the list of approved safeguards not be considered a checklist of safeguards that are appropriate in all cases. Rather, section 105 should acknowledge that different safeguards may be appropriate in different contexts and require a systematic analysis and calibration of the safeguards to the contexts, uses, threats, harms, and vulnerabilities associated with a particular research activity. In Sections 3 and 4 above, we detail a framework that can be used to develop guidance to help IRBs and investigators make such determinations.

If one of the privacy or security measures on the list of approved safeguards is found to have a serious privacy risk during the period between regular meetings of the advisory committee, a designated officer should be authorized to issue a temporary moratorium on the use of that measure until the advisory committee has an opportunity to convene. IRBs should also be authorized to approve alternate protection mechanisms, in addition to those on the list of approved safeguards, for individual studies, relying on guidance from the Secretary and the advisory committee regarding privacy risks (including a list of measures that should be considered unsafe for use). More generally, IRBs should be empowered to augment the guidance with their own guidelines regarding the selection and implementation of controls, taking into account their own expertise and institutional context. IRB decisions about alternate privacy and security controls should be reported to the advisory committee for consideration as possible additions to the list of approved safeguards.

We also call special attention to advanced data-sharing models and emerging formal approaches to privacy. There are a number of privacy methods and data-sharing models that can provide stronger privacy protection than traditional de-identification techniques that are in wide use today. Although the following data-sharing models are used across government and industry, they are often overlooked by individual investigators in favor of traditional de-identification techniques like simple redaction of identifiers:[47]

[47] *See* Salil Vadhan et al., Comments to the Department of Health and Human Services Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket No. HHS–OPHS–2011–0005 (Oct. 26, 2011), http://privacytools.seas.harvard.edu/files/commonruleanprm.pdf.

- *Contingency tables*, which are tables providing the frequencies of the co-occurrence of two or more attributes in a sample (for example, one entry in such a table might contain the number of women, age 60-65, with diabetes in the sample);
- *Synthetic data*, or a "fake" set of data that is generated from a statistical model developed using the original set of data but containing no records about an actual person;
- *Data visualizations*, which are graphical representations of the features or statistical properties of a set of data;
- *Interactive mechanisms*, or systems through which users may submit queries about a set of data and receive corresponding results, but without providing the users with direct access to the data; and
- *Multiparty computations*, which are technologies that enable two or more parties to carry out a computation that involves both of their datasets (for example, finding out how many records are shared between the two) without requiring any party to hand over its data to another.

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.[48] 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 the Common Rule, through the proposed list of approved safeguards, encourage the use of stronger privacy measures, including measures that are compatible with formal privacy models. The advisory committee of data privacy experts we propose in Section 5 should consider adding the data-sharing models discussed above to the list of privacy measures that are deemed to satisfy the requirements for reasonable and appropriate safeguards. To support their adoption, we also recommend that the advisory committee develop detailed guidance on choosing among and implementing these methods in specific cases.

The NPRM also asks whether compliance with standards from other statutes, such as the Confidential Information Protection and Statistical Efficiency Act (CIPSEA)[49] or the Family Educational Rights and Privacy Act of 1974,[50] among others, should be deemed sufficient to satisfy the safeguards requirement under the Common Rule. We caution that these standards may not be appropriate for all contexts. In particular, such standards may not be appropriate for voluntary election by investigators from a domain that is far removed from the setting contemplated by that standard. For instance, CIPSEA governs the confidentiality procedures implemented by designated statistical agencies. An individual investigator may not be well-equipped to interpret and apply a standard designed to be interpreted by a statistician at a federal agency, or to apply the statistical disclosure limitation methods used by statisticians at these agencies. For these reasons, we recommend that voluntary election to comply with safeguards from other laws not be considered generally sufficient for an exemption to the Common Rule without further guidance. Rather, the safeguards from such laws should be considered for possible inclusion in the list to

[48] Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 1 COMMUNICATIONS OF THE ACM 86 (2011).
[49] 44 U.S.C. § 3501 note.
[50] 20 U.S.C. § 1232g.

be developed in accordance with section 105 of the proposed rules. This would enable such safeguards to be reviewed and regularly updated by an advisory committee of data privacy experts, and accompanied by guidance directing IRBs and investigators regarding the contexts in which such measures are appropriate.

Question 22. Public comment is requested on whether the protections provided by the HIPAA Rules for identifiable health information used for health care operations, public health activities, and research activities are sufficient to protect human subjects involved in such activities, and whether the current process of seeking IRB approval meaningfully adds to the protection of human subjects involved in such research studies.

Response: We caution that in some cases the protections provided by the HIPAA Rules for identifiable health information are not sufficient to protect human subjects. In general, we note in Section 2 that de-identification approaches, when used alone, are substantially limited in their ability to provide systematic protection, and may lead to unexpected disclosure. In particular, we argue that the HIPAA Privacy Rule safe harbor method for de-identification should not generally be deemed a sufficient standard for protecting human subjects when releasing data. We refer to the discussion above regarding the limitations of de-identification techniques and note that the scientific understanding of privacy has advanced significantly in the time since the safe harbor method for de-identification was formulated. We recommend that this standard should not be incorporated into the Common Rule. Rather, should there be an interest in designating the safe harbor method as a safeguard that is sufficient for the protection of human subjects, it should be included within the list of safeguards to be developed in accordance with section 105. This would enable the standard to be periodically reviewed and eventually removed from the list, if deemed necessary at a future date. We also recommend that IRBs and investigators be required to consider the suitability of additional privacy and security controls when sharing data that have been de-identified using traditional heuristic techniques.

In addition, the HIPAA Security Rule requires covered entities to apply administrative, technical, and physical controls for protecting the security of identifiable health information. While it is a good practice to require information security measures, in many cases such safeguards are not sufficient. We recommend that IRBs and investigators be required to consider implementing privacy controls in addition to security controls, following implementation guidance with the elements we propose in Sections 3 and 4 above.

Question 43. Public comment is sought on the concept of requiring such minimum safeguards and limitations on disclosure, as well as whether the requirements of the proposed §__.105 would constitute a broadening of IRB responsibilities rather than a streamlining of the implementation of responsibilities that many IRBs already adopted. If an institution does view this as an inordinate broadening of responsibilities, does the institution currently have in place alternative mechanisms for ensuring data security and participant privacy in a research context? Suggestions for alternative approaches to meeting public expectation that federally sponsored research safeguard their data and protect privacy are sought during this public comment period.

Response: As discussed above, we recommend that an advisory committee of data privacy experts develop detailed guidance to assist IRBs and investigators in selecting among and implementing appropriate controls. We believe concrete instructions developed along the lines of Figure 1 in Section 4, could be used to simplify an IRB or investigator's determination regarding which safeguards are appropriate for a particular research activity. This diagram provides a conceptualization of the sets of minimum safeguards that should be considered reasonable practice in different contexts, though additional examples and anchor points in this diagram should be developed in collaboration with privacy experts and supplemented by IRBs. Guidance along these lines produced in connection with section 105 of the proposed rules could be used to establish minimum privacy and security safeguards that provide reasonable privacy protection that is calibrated to the risks and intended uses of the data.

In addition, the following approach described in the NPRM seems like a reasonable approach for reducing the burden of IRB review, if augmented by guidance and model terms for IRB policies: "It is assumed that once institutions and investigators have established policies and procedures for compliance with the new privacy safeguards at §__.105 (and it is expected that many already have such procedures in place), that IRBs will be confident in omitting that aspect of their review of research, as it does not pose unusual privacy or security risks to subjects."[51]

Question 44. Public comment is sought regarding whether the proposed Rule's information security requirements for biological specimens and identifiable private information are highly technical and require a level of expertise not currently available to most IRBs. Do these security requirements unrealistically expand IRB responsibilities beyond current competencies?

Response: We recommend that the advisory committee of data privacy experts recommended above in Section 5 be involved in the development of guidance because an evaluation of privacy risk and the suitability of various controls is highly technical and requires a level of expertise not currently available to most IRBs or investigators. IRBs, in general, are better equipped to evaluate, with guidance, the potential sensitivity of information associated with a proposed research study. In particular, we recommend that the advisory committee develop concrete minimum standards for different contexts (such as the handling of biospecimens) as well as guidelines for determining which minimum safeguards are appropriate for a particular level of sensitivity. This approach would likely reduce the burden on IRB administrators while facilitating the adoption of safeguards that strike a reasonable balance between privacy and utility.

9. Consent & accountability

Question 61. Public comment is sought on whether broad consent to secondary research use of information and biospecimens collected for non-research purposes should be permissible without a boundary, or whether there should be a time limitation or some other type of limitation on information and biospecimens collected in the future that could be included in the broad consent as proposed in the NPRM."

[51] *NPRM* at 53,988.

Response: The literature recognizes that consent should be a continuous process throughout the information lifecycle and that consent is interrelated with an individual's ongoing awareness and control. [52] Risks, and one's understanding of risks, change as uses of information expand, personal information about an individual accumulates through repeated collection and use, advances in technology improve the learning potential from a data release, and public expectations of privacy evolve over time. The complexity of the potential harms makes it unreasonable to expect subjects to comprehend and permanently commit to a complex privacy agreement at a single point in time.

Given the impossibility of foreseeing potential future uses of one's personal information and the risks such uses might carry, it should not generally be considered appropriate to obtain broad consent for secondary research uses of information. Broad consent is particularly problematic for secondary uses of information collected for non-research purposes. It is unlikely that a subject, when required to give broad consent for use of information provided outside of the research context, would have had the opportunity to reflect on the scope of, intended uses of, and risks and potential harm from future research. There may also be a significant risk of harm from use of these types of personal information, as information from non-research sources can pose risks similar to those of information from research sources.

Also note that consent is a control that can increase transparency, but it is just one control to consider as part of the overall family of controls applied across the entire lifecycle. As shown above in Figure 1 in Section 4, persistent and broad consent may be sufficient for data that have been protected with expertly selected formal methods, or where the sensitivity of the information is low, but generally it is not sufficient absent other controls. Fair information practices, as well as the framework articulated above in Sections 3 and 4, recognize a need to allow limitation on use and to provide continuing accountability through transparency, right to inspect, correct, and auditing of storage and dissemination of the data to third parties. If the final rule will permit broad consent to future, unforeseen uses of personal information, we recommend that it require IRBs and investigators to consider the suitability of additional controls from the list of privacy and security safeguards to be provided in accordance with section 105 and follow implementation guidance with the elements we propose above in Sections 3 and 4.

Thank you for your consideration of these comments as you finalize revisions to the Common Rule.

Respectfully,

Alexandra Wood (corresponding author)
Fellow, Berkman Center for Internet & Society, Harvard University
awood@cyber.law.harvard.edu

Edo Airoldi
Associate Professor of Statistics, Harvard University

[52] *See, e.g.*, Jane Kaye et al., *Dynamic Consent: A Patient Interface for Twenty-first Century Research Networks*, 23 EUROPEAN JOURNAL OF HUMAN GENETICS 141 (2015).

Micah Altman
Director of Research, MIT Libraries
Non-resident Senior Fellow, Brookings Institution

Yves-Alexandre de Montjoye
Postdoctoral Researcher, Institute for Quantitative Social Science, Harvard University
Postdoctoral Researcher, Media Lab, Massachusetts Institute of Technology

Urs Gasser
Executive Director, Berkman Center for Internet & Society, Harvard University
Professor of Practice, Harvard Law School

David O'Brien
Senior Researcher, Berkman Center for Internet & Society, Harvard University

Salil Vadhan
Vicky Joseph Professor of Computer Science and Applied Mathematics, Harvard University