October 26, 2011

To: The Department of Health and Human Services, Office of the Secretary, and Food and Drug Administration

Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket ID number HHS-OPHS–2011–0005

From:
Salil Vadhan, Ph.D. (corresponding author)[1]
Vicky Joseph Professor of Computer Science and Applied Mathematics
Center for Research on Computation and Society
School of Engineering and Applied Sciences
Harvard University
salil@seas.harvard.edu

David Abrams, J.D., M.S.
Fellow, Berkman Center for Internet and Society
Harvard University
dabrams@law.harvard.edu

Micah Altman, Ph.D.
Senior Research Scientist, Institute for Quantitative Social Science, Harvard U.
Senior Non-Resident Fellow, The Brookings Institution
micah_altman@alumni.brown.edu

Cynthia Dwork, Ph.D.
Distinguished Scientist
Microsoft Research
dwork@microsoft.com

Paul Kominers, B.S. candidate
Massachusetts Institute of Technology
pkoms@mit.edu

Scott Duke Kominers, Ph.D.
Becker Friedman Institute for Research in Economics
University of Chicago
skominers@uchicago.edu

Harry R. Lewis, Ph.D.
Gordon McKay Professor of Computer Science
Center for Research on Computation and Society
School of Engineering and Applied Sciences
Harvard University
lewis@harvard.edu

[1] On leave from Harvard as a Visiting Researcher at Microsoft Research SVC and a Visiting Scholar at Stanford University.

Tal Moran, Ph.D.
Lecturer, School of Computer Science
The Interdisciplinary Center
Herzilya, Israel
talm@seas.harvard.edu

Guy Rothblum, Ph.D.
Researcher
Microsoft Research
rothblum@alum.mit.edu

Jonathan Ullman, Ph.D. candidate[2]
Center for Research on Computation and Society
School of Engineering and Applied Sciences
Harvard University
jullman@seas.harvard.edu

We appreciate the opportunity to comment on the Advance Notice of Proposed Rulemaking (ANPRM) on human subjects research protections that appeared in 76 Federal Register 44512 (July 26, 2011). These comments address the issues of data privacy and de-identification raised in the ANPRM. Our perspective is informed by substantial advances in privacy science that have been made in the computer science literature. Our responses also support the submission made by Latanya Sweeney at the Data Privacy Lab, which more broadly addresses the fitness and appropriateness of HIPAA and the emerging field of data privacy.

We thank Alex Blocker, Allan Friedman, Dean Gallant, Bob Gellman, Becca Goldstein, Susan Landau, Deven McGraw, Tyler Moore, Latanya Sweeney, and Jim Waldo for helpful comments and discussions. The efforts of the Harvard authors on data privacy are supported in part by a gift from Google, Inc.

[2] On leave from Harvard as an intern at Microsoft Research SVC.

Question 54: Will use of the HIPAA Privacy Rule's standards for identifiable and de-identified information, and limited data sets, facilitate the implementation of the data security and information protection provisions being considered? Are the HIPAA standards, which were designed for dealing with health information, appropriate for use in all types of research studies, including social and behavioral research? If the HIPAA standards are not appropriate for all studies, what standards would be more appropriate?

Response:
No, the HIPAA Privacy Rule's standards for identifiable and de-identified information are *not sufficient* for implementing the data security and information protection provisions being considered. The HIPAA Privacy Rule is based on an overly narrow conception of what constitutes "data" and "data sharing," and consequently precludes many approaches to data sharing that could offer both better privacy and better data utility. Rather than trying to impose a single, technological standard that will have limited applicability and limited longevity, the revised Common Rule should enable a process whereby a "safe-harbor list" of data-sharing mechanisms appropriate for different contexts can be maintained and regularly updated with the input of experts and stakeholders. We note that the limitations of the HIPAA Privacy Rule detailed below also arise in the context of health data, hence we would also recommend a similar revision of HIPAA if it were in the scope of the ANPRM (and indeed, consistency between HIPAA and the Common Rule would be desirable).

Like most privacy regulation in the U.S. and abroad, the HIPAA Privacy Rule addresses the handling of *microdata,* meaning a collection of records, each of which pertains to a single individual (or household or business).[3] Although microdata is common within and outside of health care data, there is increased interest in collecting and analyzing data sets that are not in the traditional microdata form. For example, *social network data* involves *relationships* between individuals. A "friendship" relationship or contact between two individuals on a social network does not entirely "belong" to either individual's record; the relationship can have privacy implications for both parties. While this change from data about individuals to data about pairs may seem innocuous, it makes the task of anonymization much more difficult[4] and one cannot expect standards developed for traditional microdata, like HIPAA, to apply. In addition, the HIPAA Privacy Rule was conceived for microdata that arises in a traditional healthcare billing context, where each record consists of certain "identifiers" (e.g. name, address, social security number) together with fields that are typically categorical or numerical (e.g. birthdate, billing codes), and even applying it to microdata with other kinds of fields (such as text, genomic information, or locational traces) is also known to be problematic.[5] [6] [7]

---

[3] HIPAA's focus on microdata is implicit in its reference to a singular individual in its definitions of *health information* ("any information…that relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual"), *individually identifiable health information* ("information (i) that identifies the individual; or (ii) with respect to which there is a reasonable basis to believe the information can be used to identify the individual."), and *de-identified health information* ("health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual").

[4] See, for example, L. Backstrom, C. Dwork, J. Kleinberg. "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," Proc. 16th Intl. World Wide Web Conference, 2007.

[5] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays." PLoS Genetics, 4(8):e1000167+, 2008.

[6] Clifton, C., et al. "Anonymizing Textual Data and its Impact on Utility."
http://projects.cerias.purdue.edu/TextAnon/

More importantly, microdata may be converted into forms other than microdata for sharing, and these alternative forms may be difficult to evaluate under guidelines designed for microdata. Examples include:

- **Contingency tables** are tables giving the frequencies of co-occurring attributes. For example, a 3-dimensional contingency table based on Census data for Norfolk County, Massachusetts might have an entry listing how many people in the population are female, under the age of 40, and rent their home.
- **Synthetic data** are "fake" data generated from a statistical model that has been developed using the original data set. Methods for generating synthetic data were first developed for filling in missing entries, and are now considered attractive for protecting privacy (as a synthetic dataset does not directly refer to any "real" person).[8] [9] [10]
- **Data visualizations** are graphical depictions of a dataset's features and/or statistical properties. Data visualizations are especially useful for comprehending huge amounts of data, perceiving emergent properties, identifying anomalies, understanding features at different scales, and generating hypotheses.[11]
- **Interactive mechanisms** are systems that enable users to submit queries about a dataset and receive corresponding results. The dataset is stored securely and the user is never given direct access to it, but such systems can potentially allow for very sophisticated queries. For example, the Census Bureau's online *Advanced Query System* allows users to create their own customized contingency tables.[12]
- **Multiparty computations** are electronic protocols 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) in such a way that no party needs to explicitly hand their dataset to any of the others.

Sharing data in forms such as those described above (rather than restricting to microdata format) is attractive from the perspectives of both data utility and data privacy. For data utility, the above formats often allow for researchers to obtain meaningful answers to questions that they cannot obtain using microdata that has been de-identified in the manner HIPAA requires. For example: As HIPAA requires generalizing birthdates to the year, and geography to the state level, appropriately de-identified data could not be used to answer questions like, "How many babies were born with birth defects in Dauphin County, Pennsylvania during the three months after the Three Mile Island nuclear meltdown?" Nonetheless, an (approximate) answer to this question is unlikely to violate privacy -- and could be of significant utility in public health research.

In fact, many of the above forms of data sharing can often provide much stronger privacy protections for subjects than de-identified microdata. Indeed, it is now widely recognized that robust de-identification of microdata by removing and generalizing fields is quite difficult. There have been many examples of de-

---

[7] D.L. Zimmerman, C. Pavlik , 2008. "Quantifying the Effects of Mask Metadata, Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data", Geographical Analysis 40: 52-76

[8] Rubin, D. B. "Discussion of statistical disclosure limitation," Journal of Official Statistics, 1993, 9, 461-468.

[9] Fienberg, S. E. (1994). "Conflicts between the needs for access to statistical information and demands for confidentiality", Journal of Official Statistics 10, 115–132.

[10] J. Abowd and L. Vilhuber, "How Protective are Synthetic Data," in J. Domingo-Ferrer and Y. Saygun, eds., Privacy in Statistical Databases, 2008 (Berlin: Springer-Verlag, 2008), pp. 239-246.

[11] Ware, C. (2004). Information Visualization: Perception for Design, 2nd ed. Elsevier.

[12] "Census Confidentiality and Privacy: 1790-2002" (CONMONO2), http://www.census.gov/prod/2003pubs/conmono2.pdf.

identified datasets that have been later re-identified, and this has led to a heated debate among privacy law scholars about how to balance the risks and value of data sharing in a de-identification regime.[13] [14] [15]

In contrast, all of the above (privacy-aware methods for contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations) have been successfully used to share data while protecting privacy, with no major compromises as far as we know**.** For example, synthetic data has been used by both the U.S. Census Bureau[16] [17] and the German IAB,[18] and multiparty computations have been used to aggregate data across homeless programs[19] as well as in industry.[20] Moreover, many of these forms of data sharing have been shown to be compatible with a strong new privacy guarantee known as *differential privacy.*[21] [22] Although no form of sharing is completely free of risk, it seems clear that we would want make non-microdata forms of sharing an option for researchers in cases where they offer both better privacy and better utility than HIPAA-style de-identification. The benefits are particularly clear when researchers are sharing data with the public (as they are increasingly encouraged to do[23]), for whom aggregated data may be as useful as raw, unmodified data.

Unfortunately, the HIPAA Privacy Rule provides no guidance on how to evaluate privacy protections when data is shared in non-microdata formats. Conceivably, other forms of data sharing could be covered

---

[13] Paul Ohm. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," 57 UCLA Law Review 1701 (2010).

[14] Yakowitz, Jane, "Tragedy of the Data Commons" (March 18, 2011). Harvard Journal of Law and Technology, Vol. 25, 2011. Available at SSRN: http://ssrn.com/abstract=1789749

[15] Ann Cavoukian and Khaled El Emam, "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy," Discussion Paper, Information & Privacy Commissioner, Ontario, Canada, June 2011. http://www.ipc.on.ca/English/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=1084

[16] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, L. Vilhuber. "Privacy: Theory Meets Practice on the Map. Proc. 24th International Conference on Data Engineering, 2008.

[17]See Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin and John M. Abowd. 2011. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. Center for Economic Studies Discussion Paper CES-WP-11-04; which is now in use by the Census Bureau for distribution of business establishment data through the Synthetic Longitudinal Business Database. http://www.census.gov/ces/dataproducts/synlbd/

[18] Reiter, J.P., Drechsler, J. "Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality". IAB discussion paper, Intitut f¨ur Arbeitsmarkt und Berufsforschung (IAB), Nurnberg (Institute for Employment Research, Nuremberg, Germany) (2007), http://ideas.repec.org/p/iab/iabdpa/200720.html

[19] Sweeney L. "Demonstration of a Privacy-Preserving System that Performs an Unduplicated Accounting of Services across Homeless Programs." Data Privacy Lab Working Paper 902. Pittsburgh 2007, October 2008. http://dataprivacylab.org/projects/homeless/index2.html

[20] Kerschaum, et. al, "Secure Collaborative Supply Chain Management", *IEEE Computer*, Sept 2011, 38-43.

[21] Cynthia Dwork, "A Firm Foundation for Private Data Analysis" Communications of the ACM, 2011, 1, 86-95.

[22] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, "Our Data, Ourselves: Privacy Via Distributed Noise Generation." Proc., 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2006.

[23]Since 1999, all data produced under federal grants must made available to the public through the procedures established under the Freedom of Information Act (with an exemption for information that would be an invasion of personal privacy). See Office of Management and Budget Circular A-110, http://www.whitehouse.gov/omb/circulars_a110.

by a broad interpretation of the HIPAA "statistician clause," which allows one to deem information not "individually identifiable" if an expert "determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information." However, this clause is used infrequently because it requires a time-consuming, case-by-case evaluation, with no mechanism for establishing consistent standards that can be applied in the future. Furthermore, data administrators are reluctant to risk a possible legal challenge; this leads to a strong preference for the safety of the safe harbor provision.[24]

At first, it may seem that non-microdata forms of data sharing do not carry privacy risks. By definition, such forms of data sharing involve combining information from multiple individuals, and thus seem like "aggregate" data that should be safe from a privacy perspective. However, this conclusion is not correct, and it is easy to compromise privacy even while sharing data in a form other than microdata.

To illustrate the subtleties in designing data-sharing mechanisms that preserve privacy, consider an interactive system designed to answer queries about the health care expenses of the Harvard faculty, which allows queries of the form "how many Harvard faculty satisfy X" where X is a search criterion that can involve attrributes like age, health care expenses, and department. While "how many" questions may seem relatively safe when computed over a population of 2000+ individuals, they are not. By asking the question "How many Harvard faculty are in the computer science department, were born in the U.S. in 1973, and had a hospital visit during the past year?," it is possible to find out whether one of the authors of these comments (S.V.) had a hospital visit during the past year (according to whether the answer is 0 or 1), which is clearly a privacy violation. A common "solution" to this sort of problem is to only answer queries whose answers are sufficiently large, say at least 10. But then, by asking two questions --- "how many Harvard faculty had hospital visits during the past year?" and "how many Harvard faculty, other than those in the computer science department and those born in the U.S. in 1973, had hospital visits during the past year?" --- and taking the difference of the results, we can obtain an answer to the original, privacy-compromising question. A much better solution, which defeats these attacks and even achieves differential privacy,[25] [26] is to add random noise to each answer, in order to obscure the contribution of any individual. However, even here one must be careful. If we only want to add a statistically insignificant amount of noise, then the number of questions that can be safely answered is limited by the number of subjects in the dataset, and answering more questions risks serious privacy violations.[27]

The above example focuses on interactive mechanisms, but all of the forms of data sharing discussed earlier have subtle privacy risks associated with them if not implemented carefully, and these risks have been the focus of ongoing study. Aware of the risks in contingency tables, the Census Bureau has had a sophisticated disclosure review process since 1950 to guard against the leakage of individual information

---

[24] In addition,, the phrase "to identify an individual who is a subject of the information" and HIPAA's definition of individually identifiable information as information "(i) that identifies the individual; or (ii) with respect to which there is a reasonable basis to believe the information can be used to identify the individual" suggest that even the statistician clause intended to refer to microdata.

[25] I. Dinur, K. Nissim, "Revealing Information While Preserving Privacy." Proc. 22nd Symposium on Principles of Database Systems, 2003.

[26] C. Dwork and K. Nissim, Privacy-Preserving Datamining on Vertically Partitioned Databases. Proc. 24th Annual International Cryptology Conference (CRYPTO 2004), Springer Verlag, Santa Barbara, California, USA, August 2004.

[27] I. Dinur, K. Nissim, "Revealing Information While Preserving Privacy." Proc. 22nd Symposium on Principles of Database Systems, 2003, *et sequelae*.

in the tables it releases.[28] Only recently have computer scientists and statisticians begun to quantify the disclosure risks associated with methods of synthetic data generation.[29] [30] [31] Serious re-identification risks have been found in visualizations of public health data,[32] and finding privacy-preserving solutions is a subject of active research.[33] And, while there is a long history of work on "secure multiparty computation" and its use in the literature on "privacy-preserving datamining" that allows parties to jointly compute complex functions of their datasets without revealing more than the results of the computations,[34] this leaves the question of how safe the results themselves are (due to the same examples that were given for interactive mechanisms above).

It is unrealistic to hope for a one-size-fits-all set of technical requirements for data sharing. The standards and solutions that are appropriate for one form of data in one context are typically inapplicable to the others. Solutions should be tailored to the structure of the data (e.g. standard relational microdata vs. social network data vs. text), the sensitivity of the information and potential harms of dislosure,[35] the level of consent obtained from subjects, and the intended recipients of shared data. Indeed, sharing with researchers governed by IRBs, sharing with the public, and sharing under limited data-use agreements should all be treated differently. On the other hand, a case-by-case approval process by IRBs or by expert statisticians is likely to be inefficient, time-consuming, and inconsistent.

For these reasons, we propose the establishment of an evolving "safe-harbor list" for the sharing of research data involving human subjects. Each entry in this list would specify a class of data sources (e.g. electronic health records that do not include any genomic data), a class of data-sharing methods (e.g. HIPAA-style de-identification by the removal of certain fields, or interactive mechanisms that achieve a given level of differential privacy), a class of informed consent mechanisms, and a class of potential recipients. Together, these components of an entry specify a set of contexts in which a safe harbor would apply, and case-by-case IRB review could be avoided. In the long term, one can hope for this list to be sufficiently comprehensive so that the vast majority of research projects can proceed without IRB review of informational harms. But of course not all cases will be covered, and we discuss below how to better-equip IRBs to evaluate privacy risks.

---

[28] "Census Confidentiality and Privacy: 1790-2002" (CONMONO2), http://www.census.gov/prod/2003pubs/conmono2.pdf.

[29] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, K. Talwar. "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release." Proc. 26th Symposium on Principles of Database Systems, 2007.

[30] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In Proc. 40th ACM SIGACT Symposium on Theory of Computing, pages 609-618, 2008.

[31] J. Abowd and L. Vilhuber, "How Protective are Synthetic Data," in J. Domingo-Ferrer and Y. Saygun, eds., Privacy in Statistical Databases, 2008 (Berlin: Springer-Verlag, 2008), pp. 239-246.

[32] Browstein, et. al, 2006, "No place to hide -- reverse identification of patients using published maps" N Engl J Med 2006; 355:1741-1742.

[33] A. Dasgupta, Robert R., Adaptive Privacy-Preserving Visualization Using Parallel Coordinates, Transactions on Visualization and Computer Graphics (Proceedings InfoVis), 2011.

[34] Y. Lindell and B. Pinkas. "Secure Multiparty Computation for Privacy-Preserving Data Mining." Journal of Privacy and Confidentiality, 1(1):59-98, 2009.

[35] Sensitive information includes, for example, information that could be used to discriminate, as well as information that has historically made individuals uncomfortable, or has a reasonable expectation of causing embarrassment.

We recommend that this safe-harbor list be maintained by a periodically convened task force including data privacy experts from computer science, statistics, and law, members of IRBs, and researchers who do various kinds of human-subjects research, under the purview of a body such as the National Center for Health Statistics (NCHS) or the National Institute of Standards and Technology (NIST). In the foreseeable future, the safe harbor list will need to be revised quite regularly (at least once every two years), both to accommodate contexts that were not previously anticipated and to take into account new developments in our rapidly evolving understanding of data privacy risks and countermeasures (which may lead to either additions or deletions from the safe-harbor list). Such an open, deliberative, and adaptive process is likely to foster the development of practical privacy-enhancing technologies and lead to more consistent and appropriate standards than HIPAA.

While the specific technical standards that qualify for the safe-harbor list should be left to the task force of experts as discussed above, it may be important for the Common Rule to specify the general principle that the task force should use to evaluate whether a given data-sharing method should qualify for safe harbor. For this, we propose the following:

> *"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."*

Thus, the task force should not consider whether an individual can be directly associated with a particular revealed value, which only makes sense for data shared in microdata format. Instead, the task force should consider the extent to which the revealed values depend on an individual's data, and the potential harm that may result. It is important that the task force only consider harm that results from dependence on the individual's own data, because scientific knowledge gained from the dataset as a whole (which we want to allow) can and should teach us a lot about individuals, and this knowledge may be used in ways that help or harm individuals (just as the Common Rule already says that an IRB "should not consider possible long-range effects of applying knowledge gained in the research"). The determination of what constitutes "minimal risk of harm" and what constitutes "reasonably available" are judgments that will need to be guided by the experts on the task force,[36] and will vary depending on the nature of the data and the protection mechanisms used.

IRBs will also need guidance on how to evaluate the privacy risks of data-sharing methods not covered by the safe harbor, because evaluating these risks requires significant expertise and there is a rapidly advancing and highly technical literature on data privacy that should inform their decisions. Indeed, the examples given earlier about aggregate queries illustrate that there is no automatic "safety in numbers," and as a result, an IRB should ask "would the proposed data-sharing method be protective if the study consisted of a single individual?" A negative answer is an indication that technical expertise might be needed. Finally, the task force should also provide guidance and educational materials to IRBs on how to evaluate the risks of information disclosure in cases that do not fit safe-harbor criteria. In addition to the safe-harbor list, the guidance should include a "danger list" of data-sharing methods to be eschewed because they do not provide sufficient protection.

---

[36] "Minimal risk" should, of course, be interpreted consistently with the rest of the Common Rule, namely to mean that "the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests."

Regardless of implementation details, this revision of the Common Rule should be used as an opportunity to develop a forward-looking approach to data privacy, one that builds in the flexibility to react to new problems and adopt new solutions, rather than taking the outdated approach of HIPAA, which is too narrow to address even today's range of data privacy problems.

Question 55: What mechanism should used to regularly evaluate and to recommend updates to what is considered de-identified information? Beyond the mere passage of time, should certain types of triggering events such as evolutions in technology or the development of new security risks also be used to demonstrate that it is appropriate to reevaluate what constitutes de-identified information?

Response:
As discussed in our response to Question 54, we propose the establishment of an evolving "safe-harbor list" for the sharing of research data involving human subjects. We recommend that this safe-harbor list be maintained by a periodically convened task force including data privacy experts from computer science, statistics, and law, members of IRBs, and researchers who do various kinds of human-subjects research, under the purview of a body such as the National Center for Health Statistics (NCHS) or the National Institute of Standards and Technology (NIST). We envision that this safe-harbor list will be maintained by a periodically convened task force including data privacy experts from computer science, statistics, and law, members of IRBs, and researchers who do various kinds of human-subjects research. In the foreseeable future, the safe-harbor list will need to be revised quite regularly (at least once every two years), both to accommodate contexts not previously anticipated and to take into account new developments in our rapidly evolving understanding of data privacy risks and countermeasures. Such revisions may lead to either additions or deletions from the safe-harbor list.

If one of the mechanisms on the safe-harbor list is found to have a serious privacy risk in the period between task force meetings, the committee's chairperson should be authorized to issue a temporary moratorium on the use of that mechanism until the task force has an opportunity to meet. In addition, IRBs should be authorized to approve alternate protection mechanisms (i.e. ones not on the safe-harbor list) for individual studies, using guidance and educational materials from the task force regarding privacy risks (including a "danger list" of unsafe methods). Such authorizations should be reported to the task force for consideration as possible additions to the safe-harbor list.

Question 59: *Would study subjects be sufficiently protected from informational risks if investigators are required to adhere to a strict set of data security and information protection standards modeled on the HIPAA Rules? Are such standards appropriate not just for studies involving health information, but for all types of studies, including social and behavioral research? Or might a better system employ different standards for different types of research? (We note that the HIPAA Rules would allow subjects to authorize researchers to disclose the subjects' identities, in circumstances where investigators wish to publicly recognize their subjects in published reports, and the subjects appreciate that recognition.)*

Response:
No, a uniform set of protections based on HIPAA rules would neither provide sufficient protection for all information risk, nor appropriately balance the utility of protections against their costs. The HIPAA rule establishes a single level of information security protection for all identified information. This assumes implicitly that all identified information presents the same risk of harm. A basic principle shared both by the Belmont report and modern information security best practices[37] is that protections should be calibrated to the overall risk of harm. For informational harms, the overall risk is a function of both the

[37]See, e.g., National Institute of Standards and Technology Special Publication 800-30, Natl. Inst. Stand. Technol. Spec. Publ. 800-30, July 2002, ch 3, which describes the risk management framework used by all federal agencies for the selection and specification of information security controls.

likelihood of personal information being disclosed and the sensitivity of that information. As detailed in our response to question 54, both the likelihood of disclosure and the sensitivity of disclosed data vary substantially across fields of research and forms of data.

Furthermore, the technical safeguards required by the HIPAA security rules are oriented almost exclusively toward preventing access to collections of identified data by unidentified and/or unauthorized users. Some required technical safeguards such as enterprise authentication, and common technical control such as isolation from the internet, are not consistent with interactive disclosure limitation mechanisms/multiparty secure computations, since these mechanisms are explicitly provided to enable and mediate access to identified information by remote users who are not authorized for direct access, or, in some cases, may even remain anonymous. New technical controls may be appropriate for interactive mechanisms where the complexity of implementation exceeds that of current mechanisms. And, in contrast, other technical requirements, such as emergency access procedures, are simply unnecessary for most research data.

Question 63: Given the concerns raised by some that even with the removal of the 18 HIPAA identifiers, re-identification of de-identified datasets is possible, should there be an absolute prohibition against re-identifying de-identified data?

Response:
No, there should *not* be an absolute prohibition against re-identifying deidentified data or "breaking" any other privacy-protective data analysis. It has long been recognized in the computer security community that finding flaws in existing systems is crucial to developing better security solutions in the future.[38] The same is true for data privacy. Prominent re-identifications of "de-identified" datasets --- such as a Massachusetts Group Insurance Commission medical claims dataset,[39] a Netflix movie rental dataset,[40] AOL search logs,[41] and the identification of privacy risks in the aggregation of genomic data[42] --- have been instrumental in advancing our understanding of data privacy and in preventing similar privacy breaches. Thus, banning re-identification would have the opposite effect from what is intended: while it would not stop potential attackers from compromising the privacy of subjects, it would prevent honest researchers from discovering privacy leaks and ways to prevent them.

Any restriction on re-identification should contain clear exceptions for privacy research and estimating privacy risk. However, instead of restricting re-identification, it might be preferable to regulate the inappropriate use of data, e.g. using data to intentionally harm or embarrass subjects.

Of course, performing and disclosing re-identifications, even for data privacy research, must be done with care, so as to minimize potential harm to subjects while maximizing the positive impact of the knowledge gained. For this, it may be helpful to draw upon past discussions on the ethics of computer security

---

[38] Samuelson, P. "Anticircumvention Rules: Threat to Science", *Science* 14 September 2001: Vol. 293 no. 5537, pp. 2028-2031.

[39] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. In Journal of Law. Medicine and Ethics, volume 25, 1997.

[40] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In IEEE Symposium on Research in Security and Privacy, Oakland, CA, 2008. IEEE.

[41] M. Barbarao and T. Zeller. A face is exposed for aol searcher 4417749. New York Times. August 9, 2006, Page A1.

[42] J. Felch. DNA databases blocked from the public. Los Angeles Times. August 29 2008, Page A31.

research.[43] An example solution (which may not be directly applicable to re-identification) is the procedure used by the Computer Emergency Response Team (CERT) upon receiving a report of a software vulnerability: they notify the vendor of the vulnerability and wait 45 days before disclosing the vulnerability to the public.[44]

Question 64: For research involving de-identified data, is the proposed prohibition against a researcher re-identifying such data a sufficient protection, or should there in some instances be requirements preventing the researcher from disclosing the de-identified data to, for example, third parties who might not be subject to these rules?

Response:
As discussed in our response to Question 63, we oppose a re-identification ban, and believe that such a ban will lead to weaker privacy protections for subjects in the long run.

That said, all privacy protections for shared research data in the revised Common Rule should vary depending on the class of potential recipients. Sharing with the public, sharing with researchers governed by an IRB and the Common Rule, and sharing under a limited data-use agreement all require different levels and forms of protection.

Consequently, the "safe-harbor list" envisioned in our response to Question 54 would have different protection (and consent) mechanisms tailored for different classes of recipients (and different types of data).

Question 46: Under what circumstances should unanticipated future analysis of data that were collected for a different research purpose be permitted without consent? Should consent requirements vary based on the likelihood of identifying a research subject?

Response:
Yes, consent requirements should vary based on the likelihood of identifying a research subject and, more generally, should depend on the privacy risks associated with the type of data collected and the privacy protection mechanism used. The "safe-harbor list" envisioned in our response to Question 54 would have different consent requirements depending on the privacy protection mechanism, the level of sensitivity of the data, and the potential recipients of the data.

[43] David Dittrich, Michael Bailey, Sven Dietrich. Towards Community Standards for Ethical Behavior in Computer Security Research. Stevens CS Technical Report 2009-1, 20 April 2009.

[44] http://www.cert.org/kb/vul_disclosure.html