March 8, 2014

To: The Department of Labor, Occupational Safety and Health Administration
Re: Proposed Rule: Improve Tracking of Workplace Injuries and Illnesses, Docket No. OSHA–2013–0023

From: Micah Altman, Director of Research, MIT Libraries;
Non Resident Senior Fellow, Brookings Institution
David O'Brien, Project Manager, Berkman Center for Internet & Society, Harvard U.
Alexandra Wood, Fellow, Berkman Center for Internet & Society, Harvard U.

on behalf of *Privacy Tools for Sharing Research Data Project*, Harvard U.

We appreciate the opportunity to comment on the proposed rule for the electronic submission of workplace injury and illness records that appeared in 78 Federal Register 67254 (November 8, 2013). These comments address the privacy risks associated with the proposed public disclosure of such records. Our perspective is informed by substantial advances in privacy science that have been made in the computer science literature and by recent research conducted by the members of the Privacy Tools for Sharing Research Data project at Harvard University.[1]

As a general matter, we are proponents of transparency and open access, and we support the goal of making more workplace injury and illness data publicly available. We and our colleagues have previously published on the benefits to researchers and the public of wider data access.[2] We are also acutely aware of the challenges related to confidentiality that arise when collecting, analyzing, and sharing data pertaining to individuals.[3] In our research, we have studied various

[1] The Privacy Tools for Sharing Research Data project is a National Science Foundation funded collaboration at Harvard University involving the Center for Research on Computation and Society, the Institute for Quantitative Social Science, the Berkman Center for Internet & Society, and the Data Privacy Lab. More information about the project can be found at http://privacytools.seas.harvard.edu/.

[2] See, e.g., Harvard Open Access Project, Berkman Center for Internet & Society, http://cyber.law.harvard.edu/research/hoap .
A representative list of publications related to the Privacy Tools for Sharing Research Data project is available at: http://privacytools.seas.harvard.edu/publications. King, Gary. "Replication, replication." PS: Political Science & Politics 28.03 (1995): 444-452 ; Altman, Micah, and Michael McDonald. "THE PROMISE AND PERILS OF COMPUTERS IN REDISTRICTING." Duke Journal of Constitutional Law & Public Policy 5 (2010). ; Altman, M., & Mcdonald, M. P. (2014). Public Participation GIS : The Case of Redistricting. Proceedings of the 47th Annual Hawaii International Conference on System Sciences. Computer Society Press (IEEE); Altman, M., Mann, T. E., McDonald, M. P., & Ornstein, N. J. (2010). Principles for Transparency and Public Participation in Redistricting. Brooking Institute.

[3] See, e.g., Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data," Data Privacy Lab, IQSS, Harvard University (2013); Latanya Sweeney, Akua Abu, Julia Winn, "Identifying Participants in the Human Genome Project by Name," Data Privacy Lab, IQSS, Harvard University (2013); Amitai Ziv, "Israel's 'Anonymous' Statistics Surveys Aren't So Anonymous," *Haaretz* (Jan. 7, 2013),

approaches to sharing sensitive data, and we strongly believe that a sophisticated approach to data disclosure is needed in order to ensure both privacy and utility. As numerous reports by the National Research Council have shown, naïve treatment of information confidentiality and security has become a major stumbling block to efficient access to and use of research data.[4]

These comments respond to the following question raised in the Proposed Rule: "What additional steps, if any, should the Agency take to protect employee privacy interests?" We argue that OSHA's approach towards "personally identifying information" is too narrow and that releasing microdata (individual-level data) records exposes employees to re-identification risks. By citing examples from the literature, we illustrate how an employee might be identified within microdata records, even after personally identifying information, such as name, address, date of birth, and gender, has been removed. After identifying some of the risks associated with the proposed approach, we discuss data sharing models that provide stronger privacy protections and make recommendations for a more nuanced approach to releasing workplace injury and illness records.

Benefits from public data availability

Transparency is a fundamental principle of democratic governance, and public access should be the "default setting" for information collected by government. Regular reporting and wider dissemination of workplace injury and illness records will require only marginal additional effort because businesses already compile these records under OSHA regulations, this information is considered a public record, and it can easily be submitted electronically.

Routinizing the collection and dissemination of injury data is also better economic policy. It is well established that when the cost of monitoring incidents is low, more regular monitoring and gradual sanctions increase social welfare benefits.[5] Furthermore, the costs of occupational injury in the United States are at minimum tens of billions of dollars annually.[6] Most of these costs are not borne by the firms in which injuries occur, or by insurers, but are instead imposed on the

---

http://www.haaretz.com/news/national/israel-s-anonymous-statistics-surveys-aren-t-so-anonymous-1.492256

[4] National Research Council, *Expanding Access to Research Data: Reconciling Risks and Opportunities* (Washington: National Academies Press, 2005); National Research Council, *Putting People on the Map: Protecting Confidentiality with Linked Socialspatial Data* (Washington: National Academies Press, 2007); National Research Council, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research* (Washington: National Academies Press, 2009); National Research Council, *Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens and Biodata* (Washington: National Academies Press, 2010).

[5] A. Mitchell Polinsky and Steven Shavell, "The Economic Theory of Public Enforcement of Law," *Journal of Economic Literature* 38 (2000): 45-76; A. Mitchell Polinsky and Steven Shavell, "The Theory of Public Enforcement of Law, in vol. 1 of *Handbook of Law and Economics*, ed. A. Mitchell Polinsky and Steven Shavell (Amsterdam: Elsevier, 2007).

[6] J. Paul Leigh, "Economic Burden of Occupational Injury and Illness in the United States," *Milbank Quarterly* 89.4 (2011): 728-772.

individual and on the societal safety net.[7] For these reasons, reductions in injury brought about through better detection and changes in individual and firm behavior have the potential to yield substantial benefits to individuals and to the economy.

Data collected from people and institutions have proven to be increasingly useful in unexpected ways. In the area of public health, Google Flu Trends, which analyzes routine Google queries in order to provide a useful and timely supplement to conventional flu tracking methods, is a widely publicized example of the unexpected uses of data.[8] More generally, scientists are rapidly developing methods capable of mining and modeling new data sources and "big" data. Thus wider access to data and new forms of data collection are rapidly changing the study of humans, human behavior, and human institutions, and these changes are sparking scientific progress.[9]

We argue that workplace injury and illness records should be made more widely available; however, the records released to the public should not include information that is potentially private and sensitive without additional safeguards. OSHA has a responsibility to apply best practices to manage and mitigate potential individuals harms that might arise from its data releases.

Scope of information made public

If released, the information that OSHA intends to make public under the Proposed Rule could expose individuals to substantial re-identification risks. The legal standards cited by OSHA are insufficient for gauging privacy risks and protecting employees against re-identification in publicly-released records. The Proposed Rule may also prohibit the release of useful data associated with relatively low re-identification risks. Thus, the approach fails to systematically assess and balance re-identification risk, sensitivity, and utility of the data.

The Proposed Rule states that "OSHA intends to make public all of the collected data that neither [the Freedom of Information Act (FOIA)], the Privacy Act, nor specific Part 1904 prohibit from release."[10] According to OSHA, "all data fields in [Form 300A (summary log)]," "except for Column B (the employee's name), all fields . . . on [Form 300 (the log)]," and "all fields on the

[7] Id.

[8] Justin R. Ortiz, et al., "Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends," *PloS One* 6.4 (2011), doi:10.1371/journal.pone.0018687; Nick Wilson, et al., "Interpreting Google Flu Trends Data for Pandemic H1N1 Influenza: The New Zealand Experience," *Eurosurveillance* 14.44 (2009): 429-433; Samantha Cook, et al., "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic," *PloS One* 6.8 (2011), doi:10.1371/journal.pone.0023610.

[9] Micah Altman and Kenneth Rogerson, "Open Research Questions on Information and Technology in Global and Domestic Politics – Beyond 'E-,'" *PS Political Science and Politics* 41.4 (2008): 1-8; David Lazer, et al., "Life in the Network: The Coming Age of Computational Social Science," *Science* 323.5915 (2009): 721; Gary King, "Ensuring the Data-rich Future of the Social Sciences," *Science* 331.6018 (2011): 719-721.

[10] Improve Tracking of Workplace Injuries and Illnesses, 78 Fed. Reg. 67254, 67263 (proposed November 8, 2013).

right-hand side of [Form 301 (incident report)] (items 10 through 18)" could be made publicly available.[11] This excludes, according to OSHA, personal information such as name, address, date of birth, and gender from its publication requirements, but allows the publication of individual-level information such as the employee's job title, the date and time of the incident, and descriptions of the injury or illness and where and how it occurred. Although neither FOIA nor the Privacy Act likely prohibits the release of the information described above, it is not clear that these laws are suitable benchmarks for determining the scope of information appropriate for public disclosure.

The Privacy Act applies to "system[s] of records . . . from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual."[12] Since the database described in the Proposed Rule is based on establishments, not individuals, any standards or conditions for information disclosure in the Privacy Act are irrelevant since they are not readily applicable to the information in the Proposed Rule. However, if the Privacy Act was deemed to apply to OSHA's proposed database, it is not clear that OSHA could release records to the public without written authorization from the employees as the conditions of disclosure are limited.[13]

Similarly, FOIA seems to be of limited use as a standard to protect privacy in this instance.[14] Under FOIA, U.S. agencies are required to release information in response to a request, unless the agency determines that one of nine statutory exemptions applies.[15] Determinations on whether a particular exemption applies to information sought in a FOIA request are made on a case-by-case basis. Although one FOIA exemption seems on point – it exempts the release of information that "would constitute a clearly unwarranted violation of personal privacy"[16] – FOIA provides no guidance for determining the circumstances that might qualify as a "clearly unwarranted invasion." Agencies have wide discretion in such determinations. More troubling, however, is that FOIA does not explicitly mandate that information subject to an exemption *cannot* be released; rather, these exemptions are discretionary and may be waived by the agency unless another law would prohibit release. In other words, OSHA appears to have substantial deference in determining the scope of information that can be released in response to a FOIA request. FOIA is also problematic as a standard because it is architected as a request-response

[11] Id. at 67259-60.
[12] 5 U.S.C. § 552a(a)(5).
[13] 5 U.S.C. § 552a(b).
[14] Legal scholars have also raised a number of concerns regarding the privacy protections for individuals in FOIA. See, e.g., Evan M. Stone, "The Invasion of Privacy Act: The disclosure of my information in your government file," Widener Law Review, vol. 19, p. 345 (2013); Lisa Chinai, "Picture Imperfect: Mug shot disclosures and the Freedom of Information Act," *Seton Hall Circuit Review*, vol. 9, p. 135 (Spring 2013); Ira Bloom, "Freedom of Information Laws in the Digital Age: The death knell of informational privacy," *Richmond Journal of Law & Technology*, vol. 12, p. 9 (Spring 2006).
[15] 5 U.S.C. § 552(b).
[16] 5 U.S.C. § 552(b)(6); *State v. Washington Post Co.*, 456 U.S. 595 (1982).

system in which requests are individually reviewed – not for a system in which unstructured information in free form text fields is categorically released to the public. The Proposed Rule does not indicate whether similar review mechanisms would be available to weigh privacy risks against the benefits of public disclosure before or after the release of information (e.g., fields 10-18 in Form 301), which seems problematic since much of the information is descriptive yet unstructured.

The other restrictions on release in Part 1904 also seem an insufficient standard for protecting employees privacy interests under the Proposed Rule. Employers are required to omit employee names and other descriptive information that might be used to re-identify an employee only in certain types of incidents or illnesses. These are known as "privacy concern cases."[17] The scope of this restriction is limited to the types of cases specified in the regulation that involve injuries and illnesses of a sensitive character, such as "an injury or illness to an an intimate part of the body or the reproductive system," "sexual assault," "mental illness," "HIV infection, hepatitis, or tuberculosis."[18] Outside of these special privacy cases, employers are not directed by the regulations to limit the information in a way that would safeguard against re-identification, which seems underinclusive for purposes of the Proposed Rule.

As discussed in the sections that follow, many examples in current literature suggest that the combination of information that would be released under the Proposed Rule is likely uniquely identifying for many employees in the data. In fact, some individual entries for a field, such as a job title held by only one person at a company or a description of an unusual injury, may be identifying on their own. The Proposed Rule's standard for releasing or suppressing certain categories of information is weak compared to other federal regulations,[19] and there is no indication that OSHA intends to provide additional safeguards for the information that is publicly released. It also does not propose restrictions – technical, legal, or otherwise – on how

[17] 29 C.F.R. §§ 1904.29, 1904.35.

[18] 29 C.F.R. § 1904.29(b).

[19] Although the proposed data disclosures are likely not governed by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) or the Health Insurance Portability and Accountability Act (HIPAA), it is worth noting that these laws rely on definitions of personally identifying information that are significantly more expansive than the approach from the Proposed Rule. CIPSEA guidance states that "confidential information refers to any *identifiable* information, regardless of whether direct identifiers such as name and/or address have been removed from the individual records." Office of Management and Budget, *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)* (Washington: Office of Management and Budget, 2006), available at http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/proposed_cispea_guidance.pdf. In addition, the HIPAA Privacy Rule states that individually identifiable health information is 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" and that "identifies the individual or with respect to which there is a reasonable basis to believe the information can be used to identify the individual." 45 C.F.R. § 160.103. Either of these standards would prohibit the release of information such as a job title or an injury or illness description, to use just the examples above, that could reasonably be tied to an individual.

the records containing these types of uniquely identifying information may be used by the public.[20]

At the same time, the Proposed Rule calls for the routine redaction of information that, by itself, is unlikely to pose a heightened re-identification risk. For example, checkbox items 8 and 9 on Form 301, indicating whether an injury resulted in an overnight hospital stay or emergency room visit, provide information about the severity of the injury. These fields are less likely to be identifying than other fields that are proposed to be released. This redaction is substantively sub-optimal since it reduces the utility of the released data for scientific and policy analysis. It is also an indication that the Proposed Rule's classification of fields as personally identifying or non-personally identifying is arbitrary.

Re-identification risks

The Proposed Rule's requirement that certain fields be removed prior to disclosure is inadequate as a de-identification standard based on the current literature on re-identification science. It is now widely recognized that robust de-identification of microdata by simply removing and generalizing fields is quite difficult. Guaranteeing privacy protections when releasing open-ended textual data, such as fields 14 to 17 on Form 301, is especially challenging. The risks associated with de-identified microdata have led to a heated debate among privacy law scholars about how to balance the risks, sensitivity, and utility of data when sharing it with third parties.[21]

There are many examples of datasets that were believed to have been de-identified but were later re-identified. For instance, Latanya Sweeney recently demonstrated how news stories and public records can be used to identify patients in hospital records that have been anonymized by removing patient names and addresses.[22] Using information such as name, age, residence, gender, hospital, incident date, and details about the incident from news stories, as well as date of birth and ZIP code from online public records, she was able to successfully identify a significant number of individual hospitalization records in state hospital data. As another example, Kobbi Nissim and Eran Tromer recovered records of over one thousand individuals in an anonymized survey by querying an internet access mechanism hosted by the Israel Central Bureau of Statistics web site and, furthermore, demonstrated that it is possible to link these records to

[20] For example, a system that restricts access to the most sensitive data to only trusted users through technical means coupled with legal contracts specifying additional conditions on use (e.g., re-sharing of data, publishing identifying information, etc).

[21] *See, e.g.,* Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," 57 *U.C.L.A. Law Review* 57 (2010): 1701; Jane Yakowitz, "Tragedy of the Data Commons," *Harvard Journal Law & Technology* 25 (2011): 1; Ann Cavoukian and Khaled El Emam, "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy," *Information & Privacy Commissioner* (June 2011).

[22] Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data" (July 2013), available at http://dataprivacylab.org/projects/wa/1089-1.pdf.

individuals.[23]

The above examples from the re-identification literature illustrate that it is often possible to identify individuals in a data set even after fields such as name, address, gender, and date of birth have been removed. In fact, OSHA regulations also note that the descriptions of injuries and illnesses may be identifiable on their own and encourage employers to exercise discretion in describing the injury or illness in a privacy concern case if they "have a reasonable basis to believe that information describing the privacy concern case may be personally identifiable even though the employee's name has been omitted."[24] However, the Proposed Rule would require OSHA to release information that, in combination, would likely be uniquely identifying for many of the individuals in the database. This means a friend, family member, colleague, prospective employer, or even a marketer could potentially use personal knowledge of a workplace incident or information from a news article to re-identify an individual in the OSHA database and uncover sensitive details about the extent of her injury or illness.

Making workplace injury and illness records available while also providing stronger privacy protections for employees will require a sophisticated approach to data sharing. Factors such as re-identification risk, sensitivity, and utility of the data must be carefully considered and balanced when designing a system for publicly sharing personal data.

Information sensitivity

Sensitivity of the information to be released is one of the factors that should be weighed against the value and utility of the data in an information release. Generally, regulation should treat information as sensitive when that information, if linked to a person, is likely to cause substantial harm. There is a broad range of informational harms that are recognized by regulation and by researchers and practitioners in the behavioral, medical, and social science fields.[25] Types of potential harms include loss of insurability, loss of employability, criminal liability, psychological harm, social harm to a vulnerable group (e.g. stereotyping), loss of reputation, emotional harm, and dignitary harm.

Although much of the information that would be released in accordance with the Proposed Rule would likely be benign, there are some situations in which details regarding an injury or illness may be sensitive. OSHA regulations already provide additional protection for "privacy concern cases," which include injuries or illnesses related to sexual assault, mental health, or infectious

[23] Amitai Ziv, "Israel's 'Anonymous' Statistics Surveys Aren't So Anonymous," *Haaretz* (Jan. 7, 2013), http://www.haaretz.com/news/national/israel-s-anonymous-statistics-surveys-aren-t-so-anonymous-1.492256.
[24] 29 C.F.R. § 1904.29.
[25] Elizabeth A. Bankert and Robert J. Andur, *Institutional Review Board: Management and Function* (Boston: Jones and Bartlett, 2006); Raymond M. Lee, *Doing Research on Sensitive Topics* (London: SAGE, 1993).

diseases.[26] However, OSHA's exhaustive list of privacy concern cases does not include all categories of sensitive information. There are additional types of cases that involve sensitive issues, such as drug and alcohol abuse, and the public disclosure of such information would create substantial privacy risks and potential harms for the individuals involved.

Review, reporting, and information accountability

The Proposed Rule also appears to lack review mechanisms, such as case-by-base redactions of personal information that is particularly sensitive or vulnerable to re-identification risks. OSHA would not have to look far to find examples of such review mechanisms. Current OSHA regulations require employers to review and remove "personally identifying information" before sharing workplace injury and illness records with non-governmental or contracted third parties.[27] OSHA regulations also instruct employers to exercise discretion in describing the injuries and illnesses in privacy concern cases.[28]

OSHA does not seem to require similar review mechanisms before the release of this information under the Proposed Rule. The Proposed Rule calls for the public release of information such as the time, date, and circumstances of injury, even though these descriptions alone may, in some cases, be personally identifying. However, it does not indicate how, if at all, OSHA plans to systematically review and redact personally identifying information from the descriptions in these fields, or to prevent private information from being easily inferred by such redactions. OSHA may not even be aware of the extent to which identifying information might be a problem in the descriptive fields, given that it does not routinely access or collect Forms 300 and 301 outside of the limited number of investigations and inspections it conducts each year.

Moreover, the Proposed Rule appears to lack mechanisms that would provide accountability for harm arising from misuse of disclosed data. Transparency, restrictions on disclosure, and accountability for misuse are all essential to achieving an optimal balance of social benefit and individual privacy protection.[29] Accountability mechanisms should enable individuals to find out where data describing them has been distributed and used, set forth penalties for misuse, and provide harmed individuals with a right of action.

---

[26] The privacy concern cases include "[a]n injury or illness to an intimate body part or the reproductive system; [a]n injury or illness resulting from a sexual assault; [m]ental illnesses; HIV infection, hepatitis, or tuberculosis; [n]eedlestick injuries and cuts from sharp objects that are contaminated with another person's blood or other potentially infectious material . . . ; and [o]ther illnesses, if the employee voluntarily requests that his or her name not be entered on the log." 1904.29(b)(7).

[27] Specifically, employers must "remove or hide the employees' names and other personally identifying information" before disclosing information on Forms 300 and 301 to third parties." 29 C.F.R. § 1904.29(b)(10).

[28] In certain "privacy concern cases" employers are encouraged to exercise discretion in "describing the injury or illness" on both the OSHA 300 and 301 forms" if they "have a reasonable basis to believe that information describing the privacy concern case may be personally identifiable even though the employee's name has been omitted." 29 C.F.R. § 1904.29.

[29] Daniel J. Weitzner, et al., "Information Accountability," *Communications of the ACM* 51.6 (2008): 82-87.

Data sharing requires a nuanced approach

Addressing privacy risks requires a sophisticated approach, and the Proposed Rule does not take advantage of advances in data privacy research or the nuances they provide in terms of dealing with different kinds of data and finely matching sensitivity to risk. Like treatment of other risks to subjects, treatment of privacy risks should be based on a scientifically informed analysis that includes the likelihood of such risks being realized, the extent and type of the harms that would result from realization of those risks, the efficacy of computational or statistical methods to mitigate risks and monitor access, and the availability of legal remedies to those harmed.[30]

A modern approach to privacy protection recognizes the following three principles:

- *The risks of informational harm are generally not a simple function of the presence or absence of specific fields, attributes, or keywords in the released set of data.* Instead, much of the potential for individual harm stems from what one can learn about individuals from the data release as a whole when linked with existing data.

- *Redaction is often neither an adequate nor appropriate practice, and releasing less information is not always a better approach to privacy.* As explained above, simple redaction of information that has been identified as sensitive is often not a guarantee of privacy protection. In addition, the act of redacting certain fields of a record may reveal the fact that a record contains sensitive information, and this fact is often sensitive by itself or may lead to the deduction of sensitive information.[31] Redaction may also reduce the usefulness of the information, and alternative data-sharing methods, such as those below, can often enable analyses that would be impossible with redacted data (while providing stronger privacy protections).

- *Thoughtful analysis with expert consultation is necessary in order to evaluate the sensitivity of the data collected and their associated re-identification risks and to design useful and safe release mechanisms.* Naïve use of any data sharing model, including those we describe below, is unlikely to provide adequate protection.

The examples in the next section describe how these three principles can be incorporated in an information release that represents a modern, nuanced approach to privacy-protecting data sharing.

[30] Salil Vadhan, et al., *Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections* (2010), available at http://dataprivacylab.org/projects/irb/Vadhan.pdf.
[31] See Krishnaram Kenthapadi, Nina Mishra, and Kobbi Nissim, "Denials Leak Information: Simulatable Auditing," *Journal of Computer and System Sciences* 79.8 (2013): 1322-1340.

Improved data sharing models

Below we provide a list of examples of data sharing models that can be used in conjunction with new privacy-protecting methods to provide stronger privacy protections for subjects than de-identified microdata. The 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).[32]
- *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.[33]

- *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.[34]

All of the models outlined above – privacy-aware methods for contingency tables, synthetic data, data visualizations, interactive mechanisms – have been successfully used to share data while protecting privacy, with no major compromises as far as we know. Each of these mechanisms can be implemented with privacy protections and, when made privacy-aware in an appropriate way, can provide strong protection. Conversely, each of these methods, used naïvely, can create privacy risks. Many of these forms of data sharing have even been shown to be compatible with a strong new privacy guarantee known as differential privacy.[35]

---

[32] Donald B. Rubin, "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics* 9 (1993), at 461-68; Stephen E. Fienberg, "Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality," *Journal of Official Statistics* 10 (1994): 115-32; John M. Abowd and Lars Vilhuber, "How Protective Are Synthetic Data?," *Privacy in Statistical Databases* (2008): 239-49.
[33] Colin Ware, *Information Visualization: Perception for Design* (Boston: Morgan Kaufmann, 2004).
[34] U.S. Census Bureau, *Census Confidentiality and Privacy: 1790-2002* (2004), available at http://www.census.gov/prod/2003pubs/conmono2.pdf.
[35] Cynthia Dwork, "A Firm Foundation for Private Data Analysis," *Communications of the ACM* (2011): 1, 86-95.

It is clear that we would want to make some of the above forms of sharing an option for members of the public and for researchers when they would offer both better privacy and better utility than de-identified microdata. Although no form of sharing is completely free of risk, tiered access is required to ensure privacy and utility for different types of uses.[36] It is likely that publishing workplace injury and illness data using multiple levels of access would bring gains in both privacy and utility.

Generally, for data that is made available to the public without significant restriction we recommend that the data release process and method should ensure that no individual incur more than a minimal risk of harm from the use of his or her data, even when those values are combined with other data that may be reasonably available.

On this end of the privacy/utility spectrum, the unrestricted public release of data might be limited to aggregate information. Such a release could be similar in detail to the aggregated information currently provided by OSHA but cover all of the firms that would be required to submit under the proposed changes. Further these kinds of aggregate statistics, can be released with both formal guarantees of privacy and accuracy using existing differentially private methods.[37] The fact that large companies will likely have large numbers of incidents means that the added noise won't harm the accuracy of the statistics by much.

We speculate that many member of the public would be likely to find that a series of contingency tables and visualizations could simplify their review and comparison of the workplace safety records of various employers. Within such aggregated releases, coding open-ended fields such as injury and illness descriptions could additionally reduce the risk that sensitive details about an individual's injury or illness will be revealed.

An intermediate level of access could be set up to enable interactive analysis of the data, through a privacy-aware model server. This server would ensure that the results provided by the analysis leak minimal private information. It could also be used to permit audits of access and to impose some click-through data use agreements providing individuals with additional legal protections from misuse.

At the same time, for a user to gain the full utility of the data, she must have rich access to information that is minimally redacted and at the finest level of granularity obtainable. In cases where such access is needed, it should be provided through a protected and monitored data environment, such as a virtual (remote-access) data enclave,[38] and complemented with legal

[36] See National Research Council reports, supra note 4.
[37] Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith: Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006: 265-284
[38] Julia Lane and Stephanie Shipp, "Using a Remote Access Data Enclave for Data Dissemination," *International Journal of Digital Curation* 2.1 (2008): 128-134.

agreements providing information accountability and appropriate restrictions on use and sharing of the data.

Finally, OSHA should consider developing a toolkit or educational materials to help employers identify information that poses a re-identification risk in their workplace records, especially if OSHA expect that its recordkeeping forms will continue to elicit textual descriptions of injuries and illnesses in the future. Such materials could help mitigate the risk that employers will include identifying information in the forms.

Conclusion

We argue that workplace injury and illness records should be made more widely available because releasing these data has substantial potential individual, research, policy, and economic benefits. However, OSHA has a responsibility to apply best practices to manage data privacy and mitigate potential harms to individuals that might arise from data release.

The complexity, detail, richness, and emerging uses for data create significant uncertainties about the ability of traditional 'anonymization' and redaction methods and standards alone to protect the confidentiality of individuals. Generally, one size does not fit all, and tiered modes of access – including public access to privacy-protected data and vetted access to the full data collected – should be provided.

Such access requires thoughtful analysis with expert consultation to evaluate the sensitivity of the data collected and risks of re-identification and to design useful and safe release mechanisms.