wiley

NEWSLETTER

Big Data, Privacy, Research, and De-Identification

December 2015

We are living in a new and challenging era of “big data,” where an increasing volume of information is being gathered about a growing range of activities and in numerous settings where data has never been captured before. Much of this data can be “personal,” or at least can be linked to other data about a person, and that’s what creates the privacy challenge and information opportunity.

We will see in 2016 a continuing and evolving debate about the principles related to this Big Data. We likely will see a revised European Data Protection Directive, which will address “anonymous” data and develop legal requirements for data that is not obviously about a person, but which the European Union regulators believe may be linked to individuals in potentially troubling ways (from their perspective).

In the United States, where there is no “general” data privacy law, the challenge for companies (and regulators and policymakers) is how to build best practices for the use and disclosure of “big data” information, while maintaining consistency with applicable laws, an appropriate public position concerning data practices, and an eye toward the increasing likelihood of some new kind of law or regulation addressing some or all of the big data issues. There is a real need for companies to think about their strategy and business opportunities in this area, with a corresponding need to understand both the current regulatory environment and the likely changes in the near future.

At the same time, while these issues present ongoing and creative challenges, we also are seeing specific developments in specific areas that may impact this issue on a broader scale going forward. Two areas that will move forward in 2016 are of particular interest—the 21st Century Cures legislation that has passed the U.S. House of Representatives and is slowly moving forward in the Senate, and the proposed (and substantial) revisions to the Common Rule governing much of the “research” that is conducted in the United States. Both of these proposals present substantial challenges and can encourage or impede innovation and opportunity depending on how they are resolved. The core challenge—for these specific issues and for the big data debate generally—is to develop appropriate privacy protections while still permitting and encouraging appropriate and beneficial opportunities from this data.

21st Century Cures

The 21st Century Cures legislation is designed to improve "innovation" in the health care industry in the United States. The massive bill covers a broad range of changes to the drug development regulatory structure and the oversight activities of the Food and Drug Administration, along with a variety of important modifications to the health care regulatory structure designed in general to stimulate health care innovation. A small piece of the legislation addresses privacy issues under the Health Insurance Portability and Accountability Act (HIPAA) privacy regulations. These privacy revisions likely are not necessary for many of the bill's larger provisions to succeed, and, as currently written, create potentially significant privacy issues without providing substantial additional benefits.

While some of the bill's HIPAA-related provisions are useful without creating significant privacy concerns (expanding certain authorization provisions and permitting off-site access to data to develop research proposals in certain situations), there are two provisions that have not received significant attention and that can lead to material privacy risks.

First, the legislation looks at revising how data can be disclosed in general for research purposes under the "health care operations" provisions of the HIPAA Privacy Rule. The current HIPAA rules appear to distinguish between permitted data analysis for internal purposes, and disclosure of research results from this data analysis. By restricting use and disclosure of protected health information (PHI) when a covered entity is "[c] onducting quality assessment and improvement activities, including outcomes evaluation and development of clinical guidelines" if the "primary purpose" of the activity is "generalizable knowledge," the current rules—without any real explanation—seem to impede communication of research results. In practice, covered entities have been conservative in their view of this language, even though the rule may give more flexibility than current activities seem to indicate (for example, if internal data analysis leads to results that may be worth publishing, this publication likely was not the "primary purpose" of the initial data analysis).

While removal of this "generalizable knowledge" limitation makes sense, the proposal goes too far. The legislative "solution" to this problem to date has been to revise or clarify the Rule to allow the use and disclosure of PHI by a covered entity for research purposes, including studies whose purpose is to obtain generalizable knowledge, to be treated as the use and disclosure of such information for "health care operations." This would turn all "research," which today is governed by particular privacy rules for research studies, into "health care operations" where use and disclosure is largely unrestricted. This "opens up" data for a broad variety of research projects without the current controls that are in place for these projects. This may "solve" some of the "problem" (although it may not get more data to true "researchers"), but also opens up a wide variety of additional concerns. Privacy advocates and responsible data users need to be paying more attention to this issue.

Second, the House bill also raises issues related to how certain kinds of entities involved in research essentially can pay for access to health care data. While the legislative language is very hard to follow, the language appears to permit disclosure of any PHI to pharmaceutical companies, for their research, without any particular controls, and would at the same time permit these companies to pay unlimited amounts for these data. If this is the intent of this legislation, it goes way too far (by permitting unrestrained disclosures to pharmaceutical companies without any new controls), and raises enormous red flags by permitting unlimited payments for these data (at least under the HIPAA rules—other health care laws also may come into play). If the proposal were streamlined in important ways—such as by limiting the scope only to limited data sets, with accompanying data use agreements—the proposal might be worth additional consideration. Today, however, the language creates privacy risks, and there has been little attention paid to these provisions in the course of the debate about the broader range of topics being addressed.

The Common Rule Proposals

The proposed revisions to the "Common Rule" governing human subjects research also focus on this innovation idea, with the goal of streamlining the regulatory process related to human subjects research, to simplify the process without increasing reasonable harms for patients and other individuals. The proposed rule, developed by the U.S. Department of Health and Human Services and 15 other federal departments and agencies, has been several years in the making, and likely will be reviewed throughout 2016 (with formal comments due in early January 2016).

Unlike the 21st Century Cures legislation, where the privacy issues take a back seat in the public debate to other topics, privacy is front and center in the Common Rule proposal. The proposal is focused on research requirements and reducing administrative burdens where these steps can be streamlined without adversely impacting individual privacy. In general, it is important to simplify the overall provisions of the Common Rule so that research can proceed more efficiently with appropriate restrictions and protections for individuals in those circumstances where those protections make sense. While there likely will be ongoing tweaks, the proposal has in general developed an appropriate balance between the clear goals of these regulations.

For example, the draft creates categories of research where it is either clear that the Common Rule provisions do not apply, or where a more streamlined approach can be implemented consistent with the goal of protecting individuals. The goal is to make clear to research entities where they do not need to follow these rules, or where the rules do not impose significant additional requirements. The agencies have done a reasonable and appropriate job of developing these categories that will provide useful guidance to research entities going forward.

In general, this proposal represents a significant step forward, and one that will permit significant new research opportunities—and streamline administrative requirements in other situations, with an appropriate balance of research benefits and patient privacy. Companies participating in research projects will want to pay close attention to this proposal as it moves forward, as it will create risks and opportunities for a wide range of companies, in the health care industry and otherwise.

The De-Identification Debate

The third component of the ongoing debate in this area involves opportunities to "de-identify" data, so that the information is no longer reasonably linked to an individual. De-identification is a concept that generates significant discussion, both in its technical aspects and as a policy issue. Some countries—Australia for example—seem to be taking the position that essentially, data can never be de-identified, and that regardless of the removal of personal identifiers, data that was originally "personal" should remain that way for purposes of regulation. The current EU proposals don't go that far, although they lean toward this approach where there are ongoing linkages among data sets tied to a person, even if the person is not reasonably identifiable.

In many other countries and settings, however, there is a broader perspective about de-identification, one that more appropriately balances the benefits of de-identification in connection with research and public health, along with various other commercial benefits, without any meaningful impact on individual privacy. This is the general approach in the Common Rule proposals—where there are no meaningful privacy risks, the need for broader regulatory review does not exist.

The most detailed de-identification framework is spelled out in the HIPAA Privacy Rule, governing the regulation of protected health information in the United States. Under this Rule, if personal information is "de-identified" according to the regulatory process, the data is no longer identifiable and can be used and disclosed for any purpose. In order to de-identify, companies have the option of two approaches: a "safe harbor" model (not to be confused with the entirely distinct—and currently defunct—EU data transfer Safe Harbor) that requires the removal of specific identifiers, and an "expert determination" method where an expert of appropriate background and experience can review a data set and determine that there is a "low risk" of re-identification of any individual. The difficulty of this technological analysis has led to various entrepreneurial efforts to engage in sophisticated and competent de-identification, through companies like Privacy Analytics in Ottawa, Canada. We also have seen efforts by other innovative companies to address de-identification as both a "privacy protection" device and an effective information security control to protect information even when it is being held for appropriate purposes (see, for example, Anonos Inc., which has developed a patented "Dynamic De-Identification" technology to protect data and preclude privacy harms). Other companies (such as IMS Health) are thought leaders on appropriate and effective use of de-identification technologies and the public and private benefits of meaningful de-identification processes and additional safeguards.

Companies involved in the "Big Data" process will want to pay close attention to both the regulatory developments and the related technological innovations, as appropriate de-identification remains an important element in the overall goal of benefiting from data without creating undue privacy risks. Because of the volume of data generated by many companies, and the public and private opportunities that can arise from appropriate use of this data, smart de-identification has become a critical component of most companies' overall data strategy.

The Result

The ongoing discussion of big data is quite complicated, and likely will not be "resolved" anytime soon. A couple of key points are clear. The opportunities to gather data from unusual sources is clearly growing, and will continue to expand for the foreseeable future. This data will generate important and useful information to guide and improve an enormous number of activities, for both public and private benefits. And, most of this information will be useful and relevant without any need for the analysts to engage in any activity that identifies specific individuals or that creates meaningful privacy risks. So, at a minimum, data analysis using big data should be encouraged and supported, as long as there are reasonable privacy protections. These protections should include appropriate security protections (because de-identified data or other masked data is most "at risk" when there are security breaches that make this data publicly available). Opportunities to use and analyze data that has been de-identified should be encouraged, as long as reasonable de-identification practices are followed (it is clear that virtually every study that has re-identified any kind of data has been based on flawed or cursory de-identification efforts). The HIPAA standard—viewed as the gold standard of de-identification—should be encouraged as a model going forward, with the idea of appropriate controls and privacy protection steps, along with contractual and security controls, creating an environment where there is a "small risk" of re-identifying any individuals. Policymakers should examine a potential bar on re-identification (perhaps with isolated exceptions where an individual clearly would benefit from the re-identification—e.g., the data indicates a significant but undisclosed health problem in an individual). Otherwise, de-identification should be viewed as an appropriate step to get a "win-win" from big data—beneficial opportunities to improve activities, in health care and otherwise, from this broad array of personal data without meaningful privacy risk.