
PRIVATE
Whooping Cough
Other
Packs/Day
Drinks/Week
Cups/Day
Reason
MINERALS

By Alexander Gelfand

Privacy & Biomedical Research: *Building a Trust Infrastructure*

Trust. It's the basis of every patient/physician interaction: Shared personal health information is kept confidential and used only for the patient's benefit. It's a tradition that started before the time of Hippocrates, endured through the era of records stored in filing cabinets, and persists today as we move to electronic patient records. And it's codified in the form of HIPAA, the federal Health Insurance Portability and Accountability Act, which ensures the privacy of health records.

But as sensitive personal health data accrue in ever larger databases, concerns over privacy breaches are on the rise. And as researchers perceive the potential usefulness of this vast data trove, they seek strategies to access it without violating HIPAA. In response, data privacy experts are developing ever more sophisticated methods to protect electronic health data from unwanted exposure. And while many of these experts have raised alarms about the vulnerabilities of the privacy-protection schemes currently in place, they have also begun talking about the possibility of implementing far more powerful technologies in the near future.

"This is the start of the golden age of privacy research," says **Dan Kifer, PhD**, a computer scientist at Pennsylvania State University who has investigated privacy-preserving techniques for applications ranging from biomedical research to the U.S. Census.

Privacy Fears Drive Innovation

The rapid progress taking place in privacy research in the biomedical arena is driven in large part by fear—namely, fear that the vast warehouses of biomedical data now being assembled could be vulnerable to the same kinds of privacy

De-identification protocols suppress or modify bits of data that might allow an attacker to determine precisely to whom a particular record belongs.

breaches that have in recent years plagued such information aggregators as Google and Facebook.

Granted, the evidence of such breaches in the realm of medical records, clinical data, and genomic information remains slim. The most alarming incidents to date have involved simple failures of security, or of access control, such as the theft or loss of unsecured computers containing electronic medical records, rather than the unintentional leakage of sensitive information from large biomedical databases; and as yet, no one has reportedly been harmed by the unauthorized release of their biomedical information. Instead, the most impressive privacy breaches to date have been perpetrated by academic researchers who were trying to find weaknesses in the systems they were attacking: identifying the medical records of a particular individual in a hospital system, for example, or identifying participants in a genome-wide association study (GWAS) designed to link particular diseases to specific genetic variations.

Yet anxiety over the possibility of more public, and more harmful, privacy breaches continues to build, the principal concern being that insurers and employers might use biomedical data to discriminate against policyholders and employees. Such fears are not unfounded. "Insurers have historically used data to make coverage determinations," says **Deven McGraw**, director of the health privacy project at the Center for Democracy and Technology, a nonprofit public interest group in Washington, DC. **Carl Gunter, PhD**, a computer scientist at the University of Illinois who studies the health information exchanges that enable hospitals to share electronic medical records, emphasizes the "dreadful risks" posed by medical identity theft, in which one person assumes the identity of another when seeking medical care, and the medical histories of both victim and thief become dangerously entangled. And there is rising concern over the privacy risks associated with genomic data in particular. As **Brad Malin, PhD**, director of the Health Information Privacy Lab at Vanderbilt University, points out, genomic data is highly distinguishable, extremely stable, and can in certain situations be used to predict the likelihood that an individual might fall prey to this disease or that one—information that could be used to deny coverage or a job.

Some of these scenarios might seem unlikely at the present time. It's doubtful, for example, that many people outside of a university computer science department would have the technical wherewithal to pick a single individual out of the mass of statistics associated with a GWAS. But technological progress has a way of closing the gap between the improbable and the probable. "There's nothing to say that what's unreasonable now won't be unreasonable in the near future," Malin says. And even the smallest risk of a privacy violation can be enough to scare a patient away from participating in a clinical trial, or persuade an institution to withhold data from researchers due to ethical or legal concerns. According to Gunter, some health information exchanges have already prohibited the sharing of electronic medical records for research purposes. "There is such fear that we need to address it before we can make full use of this data for research purposes," says **Lucila Ohno-Machado, PhD**, founding chief of the division of bio-

"This is such a critical piece of the puzzle, that we need to address privacy concerns before we plan for other activities," says **Lucila Ohno-Machado**.

medical informatics and associate dean for informatics at the University of California at San Diego (UCSD) and principal investigator for iDash (integrating Data for Analysis, anonymization, and SHaring), a National Center for Biomedical Computing.

All of this unease—over the privacy rights of individuals, over potential discrimination, and over the chilling effect that privacy concerns can have on research—has prompted a great deal of innovation amongst the mathematicians, cryptographers, and computer scientists who are working to develop mechanisms that will allow researchers to analyze biomedical data without compromising privacy.

Data-Driven Privacy Measures

Staal Vinterbo, PhD, a computer scientist in the division of biomedical informatics at UCSD, distinguishes between two broad classes of privacy-protecting mechanisms currently under development: "data-driven" mechanisms that define privacy in terms of the data itself, and "process-driven" mechanisms that define privacy in terms of how they access the data. Both are intended to let researchers perform meaningful analyses without disclosing sensitive personal information, but they stem from very different concepts of data privacy, and they often work via very different means. The most common data-driven approaches, for example, modify raw data so that it can be released without revealing sensitive information, while most process-driven approaches leave the underlying data alone and instead build privacy protections into the algorithms they use to extract it. Data-driven mechanisms came first, but process-driven ones may offer better protection—albeit at a price.

De-identification, or anonymization, is the most commonly employed privacy measure, and one that lies very much on the data-driven side of the divide. Rather than freely sharing all of the data in a group of records, de-identification protocols suppress or modify the bits that might allow an attacker to determine precisely to whom a particular record belongs. Some of these protocols, especially the more experimental ones, can be quite sophisticated. Spectral anonymization, which Vinterbo has investigated with **Thomas Lasko, MD, PhD,** a researcher at Vanderbilt, manipulates data in mathematically complex ways so that useful correlations can be maintained for research purposes even after the information has been randomized to prevent specific individuals from being identified. Synthetic data generation, which Kifer has explored, creates new data that statistically mimics the real stuff, but shields the actual participants in the original data set. But the anonymization schemes that are most often used to de-identify electronic health records in the real world simply delete or truncate specific data fields containing identifiable information like proper names and ZIP codes.

The advantage of de-identification is that it allows analysts to examine the raw data itself, albeit in altered form, rather than running queries against it from behind a privacy-preserving interface. The disadvantage is that it does not always work.

The first and most widely publicized demonstration of the weaknesses of de-identification occurred in 1997, when **Latanya Sweeney, PhD**, linked the supposedly de-identified health records released by the Massachusetts state insurance commission to the state's voter-registration rolls and re-identified the personal medical records of then-Governor William Weld. (Sweeney, who was a graduate student at MIT at the time, is now director of the Data Privacy Lab at Harvard University.)

Staal Vinterbo, PhD, distinguishes between two broad classes of privacy-protecting mechanisms: "data-driven" ones that "perturb" or modify the data in some way so that it can be released without revealing sensitive information; and "process-driven" ones that leave the underlying data alone and instead build some kind of privacy protection into the algorithms that are used to analyze it.

Sweeney's successful re-identification attack helped prompt the adoption of the HIPAA Privacy Rule in 2000. The Privacy Rule imposes restrictions on the release of "individually identifiable health information." These federally legislated constraints on disclosure are waived, however, if the data has been de-identified by applying the so-called "safe harbor" method, which involves removing 18 identifiers, including names, dates, and Social Security numbers. Since data that has been de-identified under the safe harbor method is no longer considered to be individually identifiable, it is no longer covered by the Privacy Rule, and can be freely shared.

There is a growing sense among data privacy experts that no form of de-identification will ever be good enough to meet the highest standards of privacy protection, and that the entire approach will only grow less reliable over time.

Yet there is a growing sense among data privacy experts that no form of de-identification will ever be good enough to meet the highest standards of privacy pro-

Differential privacy is achieved by introducing some random noise into the query responses. An analyst can only see the blurry answers provided by the algorithms, never the raw data itself.

tection, and that the entire approach will only grow less reliable over time. The reason for this is simple: as more and more data is collected and stored about us all—in online databases and on social networking sites, in publicly available government repositories and elsewhere—it becomes easier and easier to launch the kind of linkage attack that allowed Sweeney to re-identify William Weld's medical records. As the computer scientists **Arvind Narayanan, PhD**, and **Vitaly Shmatikov, PhD**, noted in a 2010 article in *Communications of the ACM* (Association for Computing Machinery), "any attribute can be identifying in combination with others." In other words, no matter how many fields are deleted from an individual's record, as long as there is something left for researchers to work with, there will also be enough left for re-identification.

Moreover, because de-identification techniques have for the most part been designed to protect specific kinds of data from specific kinds of attack, they lack the flexibility needed to deal with a rapidly changing data landscape. The task of anonymization will only become harder, for example, as more and more categories of information, from handwritten clinical notes to genetic sequences, are gathered and stored in digital repositories. The *ad hoc* nature of such data-driven methods also means that they tend to lack a rigorous mathematical basis; and by their very nature they can only access a very limited amount of data. As a result, it can be difficult to quantify just how much privacy protection they truly offer. Scientists can only estimate their efficacy by trying to break them—thereby proving their limitations, but not their strengths.

This is not to say that de-identification and other data-driven approaches do not have their uses. Malin and his colleagues at Vanderbilt, for example, have for several years used de-identification strategies such as k-anonymization to help protect the privacy of electronic medical records used for research purposes. "We have de-identified more than 1.5 million medical records from the Vanderbilt University Medical Center," he says. K-anonymization works by suppressing or modifying enough data to make a certain number of records (k) in a database appear to be identical. If done appropriately, the data can still be used for research, but the individuals who provided it become lost in a crowd of look-alikes. Several data privacy experts have pointed to flaws in k-anonymization—for example, an attacker who possesses sufficient background knowledge about someone can break k-anonymity and re-identify that individual– but the risks of re-identification remain low. And like any de-identification scheme, k-anonymization allows researchers to examine the raw data.

Process-Driven Privacy Measures

Still, many scientists have begun moving away from data-driven approaches and toward more mathematically rigorous process-driven ones. "Eventually, all of us realized that this is just a never-ending cycle: find a way of perturbing the data, find weaknesses, try to fix them, find more weaknesses, try to fix them," says Kifer. While a graduate student at Cor-

"Eventually, all of us realized that this is just a never-ending cycle: find a way of perturbing the data, find weaknesses, try to fix them, find more weaknesses, try to fix them," says Kifer.

nell University, Kifer helped develop a more robust refinement to k-anonymity known as l-diversity—a refinement that was soon shown to have flaws of its own. Recognizing the apparent limitations of de-identification in general, he and his fellow researchers began seeking a different path: one that involves devising ways of querying statistical databases while providing privacy guarantees that can be expressed as mathematical and statistical statements.

The first and still the most promising of these approaches, differential privacy, was proposed in 2006 by **Cynthia Dwork, PhD**, and colleagues at Microsoft Research, Ben-Gurion University, and the Weizmann Institute of Science. Dwork recognized the cycle described by Kifer from the history of cryptography. She also knew that modern cryptographers only liberated themselves from that cycle when they developed formal, provable definitions of information security that could be quantified. So Dwork and her collaborators did the same in the realm of privacy, formulating a mathematical definition of the concept that amounts to a promise to the data subject that his life will not, in Dwork's words, "change substantially for the better or the worse as a result of a computation on the data." Precisely how that is achieved is more or less up for grabs; any solution that satisfies the basic definition, which in its true form looks more like a mathematical proof than a verbal guarantee, will necessarily be differentially private.

Scientists like Dwork and Kifer are still working out how to implement differential privacy in the real world, and few applications have moved beyond the lab. In general, however, differential privacy is achieved by writing special algorithms that sit between a statistical database and an analyst who wishes to run queries against it. If an algorithm is differentially private, then the results it produces should be essentially the same independent of whether any single person is included in the database or not. One important consequence of this is that no matter what an analyst knows—no matter what background knowledge they might possess—they still cannot learn anything more about a specific individual just because they happen to be in the database. Conversely, even if an analyst were to know everything about each individual represented in the data except for one, they still should not be able to learn much about that one remaining person. No de-identification scheme can make those kinds of guarantees.

If an algorithm is differentially private, then the results it produces should be the same regardless of whether any single record-holder is included in the database or not.

In practice, differential privacy is achieved by introducing some random noise into the query responses. For example, if an analyst were to ask how many people in a database were over 5 feet tall, and the true answer was 56, then a differentially private algorithm might grab a random variable from a probability distribution and add it to the true answer, spitting out 57 instead. The noise is the difference between the response (57) and the true answer (56). "Our choice of randomness," Dwork writes in an e-mail, "makes responses close to the truth much more likely than answers that are far from the truth (which is what we want for accuracy)."

Nonetheless, attentive readers will have noticed that differential privacy does in fact work by providing slightly inaccurate results; as Dwork says, it uses probability to introduce "a little bit of uncertainty." This has two significant consequences.

First, in a differentially private setting, an analyst can only see the blurry answers provided by the algorithms; he can never examine the raw data itself. Dwork and several colleagues are currently investigating the possibility of allowing trusted individuals to view the underlying data—a situation Dwork describes as "differential privacy with a human in the loop"—but that is still very much under development. At least for now, researchers who need to see the innards of the data sets they are working with must look elsewhere for privacy protection.

Second, the question of how much noise is enough noise, and how much noise is too much noise, is a rather thorny one. The trick is to add just enough randomness to the query answers to protect the privacy of the individuals whose records lie in the database, but not so much that an analyst can no longer learn accurate or meaningful things from a statistical perspective about the sample population they comprise. This is the price of privacy, or the trade-off between privacy and utility; and it may be the most serious challenge facing those who are trying to bring differential privacy out of the lab. "We can always find wildly inaccurate ways of computing something that ensures a given level of privacy," says Dwork. The science lies in finding ways of ensuring privacy that do not destroy utility.

"We can always find wildly inaccurate ways of computing something that ensures a given level of privacy," says Dwork. The science lies in finding ways of ensuring privacy that do not destroy utility.

As it turns out, some queries, and some databases, are more "sensitive" than others, meaning that they are more prone to leak information. As a result, they require more noise. According to Vinterbo, more fine-grained infor-

mation also requires more noise—a situation that may prove challenging in the case of genomic databases, which contain enormous amounts of incredibly detailed data. Since you wouldn't want to add more noise than is absolutely necessary, matching the appropriate amount of noise to the sensitivity of the query and of the database—in effect, figuring out how to balance privacy against utility—is crucial. And as **Kamalika Chaudhuri, PhD**, an expert on machine learning at UCSD, says, it also turns out to be "fairly technical and complicated."

Which is not to say that it can't be done. A number of researchers are investigating ways of relaxing differential privacy so that it still offers strong privacy protection without requiring excessive amounts of noise, while others are trying to find novel methods of adding noise that won't degrade accuracy.

Chaudhuri, for example, is interested in using algorithms called "classifiers" that can be used to trawl through large collections of medical records in order to predict things like whether a particular individual might require hospitalization. Classifiers must be trained on standard data sets, however, and the training process can leak sensitive information about the training samples. A differentially private approach would typically involve adding a bit of noise to the results coming out of the classifier—a technique known as "output perturbation." This protects privacy, but also makes the classifier more error-prone. Chaudhuri has figured out a way to insert the noise earlier in the process, injecting it into the classifier itself—a technique she calls "objective perturbation." The latter still ensures differential privacy, but the results are more accurate.

Efforts like these bode well for the adoption of differential privacy in the coming years. But even its supporters agree that differential privacy alone cannot be counted upon to solve the privacy problem once and for all. "There is no single solution that will suit every possible scenario," says Vinterbo. In a recent paper, Kifer pointed toward a few specific weaknesses of differential privacy, most notably some limitations on its ability to protect privacy in social networks and in circumstances where some statistics have already been released into the wild. "Differential privacy works," Kifer says. "But nothing works all the time."

"Differential privacy works," Kifer says. "But nothing works all the time."

Finding Integrated Solutions

As a result, many experts are beginning to envision a more integrative and contextual approach to biomedical data privacy—one that would offer a menu of technical solutions backed up by policy measures, the precise mixture of which would depend on the nature of the data, the needs of the researchers, and the concerns of the data subjects themselves.

Haixu Tang, PhD, and **XiaoFeng Wang, PhD**, at the Indiana University Bloomington School of Informatics and Computing, advocate for what they call a "hierarchical method of data release" that would consider the kind of analysis researchers wish to perform on a particular data set, the level of privacy risk involved, and the degree of utility required before deciding on a particular privacy mechanism. (The two recently won the 2011 Award for Outstanding Research in Privacy Enhancing Technologies for their work demonstrating that individuals could be identified in a GWAS even when the precision of the published statistics was low and some of the data were missing. They are currently investigating ways of introducing miniscule amounts of noise in order to guard against such attacks without sacrificing utility.)

Vinterbo, for his part, thinks that a comprehensive solution to the privacy problem will require a "trust infrastructure" that includes not only technical solutions, but also "legal frameworks that efficiently combine technology and law." "The needs that will be met by technical measures alone are a minority," he says.

Vinterbo thinks that a comprehensive solution to the privacy problem will require a "trust infrastructure" that includes not only technical solutions, but also "legal frameworks that have some actual teeth."

Similarly, Malin would like to see a holistic, risk-based approach that draws on the pooled expertise of technologists, legal experts, and ethics review boards, all of whom would have a say in determining how best to safeguard privacy in a particular context—whether that meant implementing the most rigorous technical scheme possible, or applying something less formal and backing it up with carefully crafted use agreements and legal sanctions. Only then, he believes, will the biomedical community have the kind of flexible, nuanced tools needed to address the challenges of protecting its data.

"We have developed great technical solutions, and more are coming down the pipeline," says Malin, echoing Kifer's prediction of a golden age. "But we have to keep the bigger picture in mind." □