Preserving temporal relations in clinical data while maintaining privacy

RECEIVED 30 September 2015
REVISED 4 January 2016
ACCEPTED 6 January 2016
PUBLISHED ONLINE FIRST 24 March 2016

George Hripcsak,[1] Parsa Mirhaji,[2] Alexander FH Low,[3] and Bradley A Malin[4,5]

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY. OXFORD UNIVERSITY PRESS

ABSTRACT

Objective Maintaining patient privacy is a challenge in large-scale observational research. To assist in reducing the risk of identifying study subjects through publicly available data, we introduce a method for obscuring date information for clinical events and patient characteristics.

Methods The method, which we call Shift and Truncate (SANT), obscures date information to any desired granularity. Shift and Truncate first assigns each patient a random shift value, such that all dates in that patient's record are shifted by that amount. Data are then truncated from the beginning and end of the data set.

Results The data set can be proven to not disclose temporal information finer than the chosen granularity. Unlike previous strategies such as a simple shift, it remains robust to frequent – even daily – updates and robust to inferring dates at the beginning and end of date-shifted data sets. Time-of-day may be retained or obscured, depending on the goal and anticipated knowledge of the data recipient.

Conclusions The method can be useful as a scientific approach for reducing re-identification risk under the Privacy Rule of the Health Insurance Portability and Accountability Act and may contribute to qualification for the Safe Harbor implementation.

RESEARCH AND APPLICATIONS

INTRODUCTION

Observational research is on the rise, with expansive, even international, data networks being created.[1–3] To ensure data reuse on a large scale, these and similar programs have had to address numerous challenges, including data standardization, data quality, and patient privacy, particularly in the situation when consent has not been solicited.

With respect to the latter issue, creating a de-identified observational data set can protect patient privacy, lessen the administrative burden for obtaining (or waiving) authorization for use of the data, and reduce the risk of fines and litigation in the event of a data breach.[4–6] However, dates, which hold information that is vital to research, present a particular challenge for de-identifying observational data. For example, in studying medication side effects, it is essential to know whether a potential side effect preceded or followed a medication. Yet, dates may be exploited to re-identify cases when relevant information, such as birth or death dates, are often publically available.[7–9] In the case of data sets that are updated over time, dates present an even greater challenge. Even if the dates are removed completely from the data set, by observing when new events appear between updated versions of a database, one can narrow down the date range for the new events.

A recognized approach to address dates in de-identified records is to assign each patient a random temporal shift (eg, 1–365 days) and consistently apply that shift to each of their dates. For example, Vanderbilt's BioVU,[10] University of Chicago,[11] IMShealth,[12] and Physionet[13] use date shifting in their de-identification models. Yet one of the challenges of generalizing basic date shifting strategies to other institutional projects is that when the original identified data set has date boundaries, such as an overall start date for the data set or the current date as the end of the data set, then shifts can be inferred by merely looking for patients with data at the end of the shifted data set. Those patients will likely be assigned a value close to the maximum possible shift. Shifting dates backward instead of forward does not avoid the problem because patients with data at the end of the shifted data set will likely have the minimum shift value. Updating the data set, which reveals information about recency, enables inference about ranges of dates and shifts associated with them, as noted above.

A variation of date shifting is to replace dates with the difference in time (ie, duration) from some reference point that is unique to each patient, such as the patient's first visit.[14,15] Durations are also vulnerable to re-identification, however. Sarpatwari et al.[6] point out that even durations can reveal temporal information via interaction with clinical study periods; similarly, duration-based data sets are subject to the same update risk as shifted dates.

In this communication, we present a method to obscure dates via a shifting process, which maintains the relative temporal relationship among events, but hides the actual date of the events to any chosen granularity. This strategy, which we call Shift and Truncate (SANT), remains robust to updates no matter how frequent (even daily). As the name suggests, it applies random date shifting with truncation periods at the beginning and end of the data set. All patient data with shifted dates within the truncation period are removed from the data set such that each patient has an equal chance of adding data to the end of their record at each update without revealing what date those data occurred.

This method can be applied in several cases. The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA)[16] provides a definition of de-identified data, which can be implemented through two strategies. The first is an expert's determination, which requires the risk of re-identification of the individual to which a record corresponds be sufficiently small. Our method may be useful for HIPAA's expert determination by obscuring dates to any desired granularity (eg, month or year) and thus reducing risk of re-identification using public records. In a similar vein, this approach may be useful in supporting recent policy recommendations from the European Medicines Agency for sharing clinical trials data.[17]

The second strategy is called Safe Harbor, which requires the removal of eighteen features. Notably, Safe Harbor states that data related to a date is considered de-identified only if they communicate

Correspondence to Dr George Hripcsak, Department of Biomedical Informatics, Columbia University Medical Center, 622 W 168th St, PH20, New York, NY 10032, USA; hripcsak@columbia.edu For numbered affiliations see end of article.

©The Author 2016. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved. For Permissions, please email: journals.permissions@oup.com

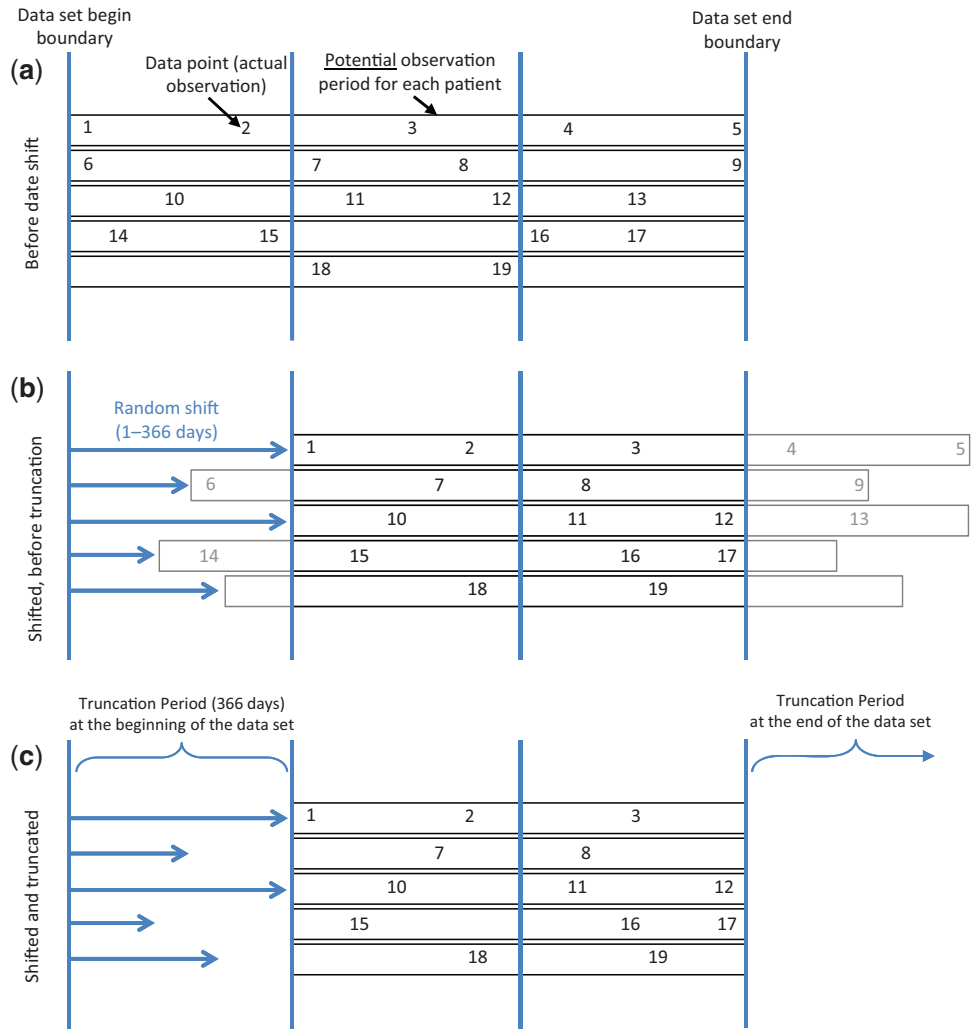
knowledge of no greater detail than 1 year in length. A naïve approach to satisfying this requirement is to simply drop the month and day of all dates. Unfortunately, doing so eviscerates essential information about ordering and timing of medical events and still reveals information finer than year if the data set is updated over time. Our proposed method can hide information finer than a year despite more frequent updates yet maintain temporal relationships.

METHODS

Following the example of Safe Harbor, let us assume that one wants to obscure the date of events to within a year. Under the SANT method, each patient is assigned a unique number randomly selected from a range of 1–366 (to ensure coverage for leap years). This number is added to each date in the patient's record, as shown in Figure 1a. A date shift by itself would reveal temporal information about patients with the latest dates in the data set because they must have had a high (near 366) shift. Events are therefore removed (the entire event, not just the date) to ensure that every patient has an equal chance of revealing new information at the end of the data set. The truncation is applied as a 366-day period at the beginning and end of the data set (for a total of about 2 years of truncation). For example, imagine the data set contains data from January 1, 2007 to January 1, 2015. Then events whose shifted dates occur before January 2, 2008 or after January 1, 2015 are removed from the data set, as shown in Figure 1b. Theorem 1 asserts that our method successfully obscures all temporal information finer than m, which we can set to 366 days.

Theorem 1: Let $X = \{x_1, x_2, .., x_n\}$ be the set of original dates in the system for a specific patient. Let $r \in \{1, 2, \ldots, m\}$ be an offset drawn

Figure 1: Shift and Truncate. Each row is a unique patient, each number is a unique data point for a patient, and each rectangle represents the time that the patient was *potentially* observed. (**a**) Original data set. Patients are potentially observed for 3 years (each vertical line marks 1 year). Patients need not have data, but simply the potential to have been observed (even if they lived elsewhere or were not born yet, someone had the potential to have been observed). (**b**) Shifted data set. Patient records are shifted forward by 1–366 days. Data points that were previously aligned across patients are no longer aligned, but points within a given patient remain at the same relative distances from each other. (**c**) Shifted and truncated data set. Data points from the first 366 days of the shifted data set and from the last 366 days of the shifted data set are removed from the data set.



from the set uniformly at random, a be the first date that data were recorded in the original system, and b be the last date that the data in the original system were known to be recorded. Let $Y=\{y_1, y_2, \ldots, y_n\}$ be the set of shifted dates, such that:

$$y_i = \begin{cases} x_i + r, & \text{if } (a+m) \leq (x_i + r) \leq b \\ \varnothing, & o.w. \end{cases},$$

where $\varnothing$ implies that the date is not reported. There exists no y_i such that the underlying x_i, can be bounded to range smaller than a size of m.

Proof: Let us assume there exists an i such that y_i is reported and x_i can be bounded to a range smaller than a window of m dates. This implies that r is bounded to a value range smaller than a size of m because there is a direct relationship between r and x_i. This is a contradiction in the definition of r and, as a consequence, it follows that no such i exists. ■

As a data set is updated, the truncation date at the end of the data set will shift forward, and events that were previously removed will then be added back to the data set. In the above example, if an update occurs three months later, then events after March 1, 2015 will be removed, but events in the January 2, 2015 to March 1, 2015 interval will be added back or added anew (ie, events for patients with a low shift value and recent data).

It is important to recognize that events in the data set that fall within the truncation window must be withheld entirely and not just have their dates set to null. This is because events that transpire in the future truncation window that are dateless today will have dates assigned in subsequent updates. By comparing a pair of updates, shifts can be narrowed down for patients. For example, imagine there is an event initially reported with a null date. After a future update, when the shifted date of this event is revealed to be m (eg, 366) days later than the original truncation point, one can infer that the patient's shift must be m. In general, the shift can be narrowed down to a range equal to $m+1$ minus the number of days the last data point exceeds the original truncation point. If the events are instead deleted from the original cohort, then one cannot infer the shift.

Birthdates are a special case that we wish to highlight. They are also shifted to maintain consistency of the data set and to avoid revealing the shift. For instance, a child born within the data set must have the birthdate and the hospital records of the delivery match. Yet birthdates need not be truncated if they fall in the one-year window of truncation at the start of the data set. Unlike clinical events, which are present only within the data set's time span, such as admissions, birthdates are recorded retrospectively and, thus, do not commence at the beginning of the data set. In fact, it would make no sense to truncate in such a scenario because most patients will have birthdates before the truncation window. Unlike birthdates, clinical events related to the birth, such as an in-hospital delivery, should be truncated like all other events, however. That is, the birthdate would remain but the hospital admission for the delivery would be removed.

When a birthdate lies at the end of the data set post-shift, then it should be truncated like all other events. This is because birthdates do not occur beyond the end of the data set; we do not predict future birthdates. In the latter case, retaining the birthdate would reveal the true birthdate for some patients due to the fact that a birthdate 366 days into the future must have occurred at the time of the update. Consider, if there were prenatal data attached to the child instead of the mother, then those data would precede the birthdate and knowledge of the true birthdate would reveal knowledge of the prenatal data, which might precede the truncation window. Truncation of birthdates at the end of the data set avoids this disclosure (in this example, the "patient" would have to have no birthdate because the birth would be in the future). Generally, the patient will have no other data in the data set, so truncation at the end does not lose additional data.

Furthermore, mother-child links, donor-recipient links, and all other inter-patient links must be deleted or the shifts for the pair must be the same random value. Otherwise the shifts may be narrowed down from the differences in dates that are applied to the same event (eg, a mother's delivery and a baby's birth date). Also, the shifts cannot be changed to accommodate a new relationship (donor-recipient). Such links provide greater risk for re-identification independent of dates anyway, so their deletion may be prudent.

The SANT procedure obscures temporal information finer than any chosen granularity despite frequent updates and yet maintains relative temporal information. It works by producing a data set in which each patient is equally likely to have new information added to the data set in each update. It should be a useful tool for de-identifying clinical data sets.

Depending on the goal, one can choose a shift of integral days or continuous time, where the latter obscures both the date and the time-of-day. Under HIPAA Safe Harbor, time-of-day is not excluded. Because time-of-day can also be useful for research, especially in studies that are dependent on patient physiology, a shift with integral days can be helpful.

It should be noted that the SANT procedure does not address the problem of inferring dates from clinical content or family structure.[18,19] Influenza is more likely during influenza season. Boating accidents are more likely over the summer. Certain drugs may be introduced (or come into vogue) at certain points in time. These and other clues are available under all forms of de-identification, however, with or without SANT. Even data with no dates whatsoever will have some risk, and data with that retain information about relative durations – including SANT – carry additional risk because finding the time of one event reveals information about others.

Like all date shifting methods, SANT preserves relative temporal information but loses information about seasonality depending on the chosen granularity. For example, with a coarse granularity like a year, syndromic surveillance will be difficult or indirect. Furthermore, Liu et al.[14] point out that date shifting can induce the side effect that occasionally diagnosis or procedure codes may become invalid if the date shift moves them from a time when they were valid to a new date when they are invalid as a terminology evolves.

SANT's truncation step loses additional information. At a maximum, patient records will be truncated on either end up to the chosen granularity (see Figure 1b), and, on average, patient records will be truncated to half the chosen granularity from each end. Recent events, such as a new pandemic, will be underrepresented to a greater degree as it approaches the end of the set. Put another way, seasonal events will be spread out over the year, and the spread may be partially truncated at either end of the data set. Data ought to be rolled back to what was known at the date of truncation. For example, if a tuberculosis culture result, which takes six weeks to grow, is obtained from a patient just before the truncation time and shows tuberculosis then one will know that the shift had to be at least 6 weeks. It should instead be rolled back to what was known at the point of truncation.

In our review of the literature, we found examples of date shifting[10–13] and use of duration[14,15] but relatively little discussion of the ability to infer shifts or true dates at the extremes of the data set or with updates, with Sarpatwari et al.,[6] noted above, being an exception.

We have found no mention of truncation as a solution to the inference of dates at extremes and with updates.

One can compare the various ways of handling dates using the following concrete example, which assumes m is 366. We will focus on the end of the data set, but the beginning is similar. Assume that the database is being updated December 31, 2014 and that that is the last day that patients may have information. Assume that a given patient, "A," has events on March 1, 2014 and on November 1, 2014, and that the patient's randomly chosen shift is 300 days. A simple date shift without truncation will shift the apparent dates to December 26, 2014 and August 28, 2015. The last possible shifted date in the data set will be January 1, 2016, and with a large enough data set, there may be an actual event for another patient, say "B," with that date. Given this information, then even in a static data set, one could surmise that patient B had a shift at or near 366 with an event at to near December 31, 2014, and that patient A, whose last date is August 28, 2015, has a shift of about 240 days or higher (ie, 366 days minus the duration between August 28, 2015 and January 1, 2016), and that the latter event most likely occurred sometime in the range of August 27, 2014 to December 31, 2014. The exact range will be dependent on the size of the data set and sparseness of the data, which will determine the likelihood of having at least one event on the very last possible day. For example, a data set with 100 000 cases where each patient has one event per year has over a 50% chance of having an event on the very last possible day (ie, $1– (1–(1/366)^2)^{100\,000}$ based on the probability of having an event on the last day and also having the maximum shift), and a data set with 1 000 000 cases would have over a 99.9% chance. Note that date shifting's vulnerability in static data sets only occurs if the creation of the data set is synchronized; if patients are added one at a time over a period of time, then the above inference cannot be done.

With SANT, the shifts would be the same, but shifted events dated beyond December 31, 2014 would be deleted, resulting in a loss of the original November 1, 2014 event (shifted to August 28, 2015) but not the March 1, 2014 event (shifted to December 26, 2014). The latest event in the data set for any patient would be December 31, 2014, and one could not infer anything about the shifts. All patients would be equally likely to have data on the last visible day, December 31, 2014. Some patients would have 366 days of data deleted from the end, and some would have a single day of data deleted from the end.

An alternate approach would be not to shift dates but to delete dates altogether or just retain the year, as nominally suggested by Safe Harbor. This loses the temporal relationships among medical events, but might appear to be a low-risk approach. While it would hide temporal information for a static data set, it would not for one that is updated (and therefore such a data set would not qualify for Safe harbor). For example, if patient A had a subsequent (third) event on January 15, 2015, and if the data set were updated monthly, say at the end of each month, then on January 31, an event would appear in the data set for patient A that was not there before, and one could surmise that it occurred some time in January even if no date were attached to it. Under SANT, that event would not appear in the data set until the update of November 30, 2015, and it would appear as a November 11, 2015 shifted event. More important, one could not infer the shift or the true date of the event beyond the fact that it had occurred some time within the 366 days before November 11, 2015.

Replacing dates with relative durations would represent the event times for patient A with respect to some initial event in A's past. For example, if patient A's first event was January 1, 2010, then the March 1, 2014 and November 1, 2014 events would be 1520 days and 1765 days, respectively. Each patient would have a potentially different origin. All year information would be lost. Similar to the date-deletion method, the duration method would hide temporal information for a static data set but would be subject to the same risk for an updated data set: the January 15, 2015 event would show up as 1840 days and by its new appearance, it could be inferred to be recent.

Perhaps the most important competing method to SANT is to retain true dates. Retaining dates clearly violates Safe Harbor but may still be statistically determined to be of low risk for re-identification. This can work for very large data sets (say, 100 000 000 patients) that have publicly known dates removed, such as births, deaths, and marriages, and that do not retain fine geographic data. The obvious benefit is the presence of real dates for medical events with the ability to study fine temporal relationships and to track seasonal effects. The challenge is that not only must the primary dates for birth, death, and other events be deleted but this must be done for related events as well. For example, not only must the birthdate be masked but also a related labor and delivery and potentially some prenatal and perinatal data. As new public data sources come on line, the data set must be rechecked for re-identification risk.

To summarize the alternatives (see Table 1), eliminating all dates or all date components finer than a year loses the ordering of medical events and only works for static data sets; frequent updates will reveal temporal information. Shifting dates without truncation maintains the ordering of medical events but loses the ability to study seasonal effects and remains vulnerable to inferring true dates in both static and updated data sets. Using durations addresses static data sets but not updated ones, and it maintains the ordering of events but not seasonal effects and it loses even coarse temporal (year) information. Retaining full dates allows the finest temporal analysis including seasonal effects, but it requires large data sets and complicated expunging of selected dates. SANT allows fine temporal analysis although without seasonal effects, it can work on smaller data sets, and it can be updated frequently. It loses data at the ends of the data set, however, about half a year from each end on average.

We point out that the methods that do not support updates will soon fall behind an updated SANT data set: events that are truncated today will eventually appear under SANT, but a static data set will fall further and further behind with new events never appearing. It may, on the other hand, be acceptable to update a data set once per year even with the methods that do not support updates under the argument that it would reveal information only to the year level. In this case, one would be behind on events at the end of the data set by one-half year on average, the same as SANT.

The question of whether the SANT procedure can qualify for HIPAA Safe Harbor depends on several interpretations. First, does the inclusion of any dates with month and day, even false dates, break the HIPAA requirement that elements of dates except year be removed for dates directly related to an individual? One way to assess this question is to consider the following scenario. If a data owner shared a data set with only years and then a recipient added purely fictional months and days to it – eg, January 1 to every data element – does the data set now fail safe harbor? Because no information would have been added to the data set – or, put another way, because the false dates are not "directly related to an individual" – one would think that the data set would not be considered re-identified and would still qualify for Safe Harbor. In addition, in many database systems, the way to truncate month and day is to set all data to an arbitrary date like January 1. We therefore believe that the inclusion of false dates may not disqualify a data set for Safe Harbor.

Second, does the inclusion of shifted dates fail the Safe Harbor provision of not using derived identifiers because dates are shifted

Table 1: Comparison of Date-obscuring Methods

Method: feature	Drop all dates	Drop month and day	Shift dates	SANT	Durations only	Retain nonpublic dates
Protects against date-based re-identification in a static data set	Y	Y	N	Y	Y	Only if large data set and expunge selected dates
Protects against date-based re-identification in an updated data set	N	N	N	Y	N	Only if large data set and expunge selected dates
Retains all available events	Y	Y	Y	N (drops 1 year of events, 1/2 year from each end on average)	Y	Y
Fine (day) ordering of medical events maintained	N	N	Y	Y	Y	Y
Seasonality maintained	N	N	N	N	N	Y
General time frame (year) maintained	N	Y	Y	Y	N	Y
Supported by standard DBMS tools	Y	Y	Y	Y	N	Y
Qualifies for Safe Harbor	Yes only if no updates	Yes only if no updates	Possibly only if no updates	Possibly (with or without updates)	Probably only if no updates	N

from (derived from) the original dates? The HIPAA provision about derived codes is specifically discussed in the context of re-identification codes and not elsewhere.[16] The shifted dates are not used (and are not able to be used) to re-identify individuals, and they are not uniquely identifying in their own right, so they may not come under the provision of derived re-identification codes. In addition, a formal verification by a statistical expert that the shifted dates cannot be used to infer temporal information finer than year may also provide argument that Safe Harbor is not violated. Using durations since a patient-specific origin such as first visit combined with truncation with year granularity to avoid date disclosure due to database updates may also qualify for Safe Harbor. Nevertheless, because there is no mechanism to verify the interpretation of HIPAA other than waiting for legal cases, it may be wise to have the data set certified as low risk of re-identification rather than rely solely on Safe Harbor. A hybrid approach, which uses durations along with SANT-like truncation might be more likely to satisfy Safe Harbor – because no "dates" would be included, only durations – at the expense of losing coarse (year) information and difficulty storing it in database management systems designed to hold dates.

SANT is currently being used by five medical centers and additional clinical collaborators as part a Patient-Centered Outcomes Research Institute Clinical Data Research Network[20] and as part of an international Observational Health Data Sciences and Informatics network.[1] The authors note that use does not imply correctness, and each potential user must assess the risks and benefits of the approach.

CONCLUSION

Our SANT method for obscuring dates finer than any chosen granularity is a tool for managing re-identification risk. Unlike more naïve approaches (like dropping month and day or simple date shifts) it can be proven not to reveal temporal information despite frequent data set updates. The method maintains relative temporal relationships, which can be critical for observational research, but it does lose absolute dates and seasonality (depending on the granularity) and it does truncate clinical data at the ends of the data set. It may be useful for reducing re-identification risk under HIPAA and may contribute to Safe Harbor qualification.

CONTRIBUTORS

All authors made substantial contributions to the conception and design of the work; drafted the work or revised it critically for important intellectual content; had final approval of the version to be published; and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This work was supported in part by grants from the National Institutes of Health, R01LM006910, R01LM009989, and UL1 TR000135.

COMPETING INTERESTS

None.

REFERENCES

1. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–578.
2. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby J, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc.* 2014;4:578–582.
3. Gottesman O, Kuivaniemi H, Tromp G, *et al.* The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genet Med.* 2013;15:761–771.

4. Cavoukian A, El Emam K. *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy.* Office of the Ontario Informational and Privacy Commissioner, Toronto, Canada. June 2011. https://www.ipc.on.ca/images/Resources/anonymization.pdf. Accessed August 31, 2015.
5. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *J Am Med Inform Assoc.* 2013;20: 29–34.
6. Sarpatwari A, Kesselheim AS, Malin B, Gagne J, Schneeweiss S. Ensuring patient privacy in data sharing for postapproval research. *N Engl J Med.* 2014;371:1644–1649.
7. Sweeney L. Weaving technology and policy together to maintain confidentiality. J Law Med Ethics. 1997;25:98–110.
8. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *J Am Med Inform Assoc.* 2010;17:169–177.
9. El Emam K, Dankar F, Vaillancourt R, Roffey T, Lysyk M. Evaluating the risk of re-identification of patients from hospital prescription records. Can J Hosp Pharm. 2009;62:307–319.
10. Roden DM, Pulley JM, Basford MA, *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008;84:362–369.
11. Human Imaging Research Office (HIRO), University of Chicago. Image de-identification and distribution specifications. https://hiro.bsd.uchicago.edu/image-specs. Accessed September 22, 2015.
12. Privacy Analytics. *IMS Health: Unlocking the Value of EMR Data for Advanced Research and Analysis, Better Health Metrics, and Product Innovation.* http://www.privacy-analytics.com/files/IMS-Brogan-Case-Study.pdf. Accessed September 22, 2015.
13. Neamatullah I, Douglass M, Lehman LH, *et al.* Automated de-identification of free-text medical records. *BMC Med Inform Decision Mak.* 2008;8:32.
14. Liu J, Erdal S, Silvey SA, *et al.* Toward a fully de-identified biomedical information warehouse. *AMIA Annu Symp Proc.* 2009;370–374.
15. QuesGen Systems. *Three Ways to Deal with HIPAA Dates in De-Identified Data Sets.* http://www.quesgen.com/three-ways-to-deal-with-hipaa-dates-in-de-identified-data-sets/. Accessed September 22, 2015.
16. U.S. Dept. of Health & Human Services. *Standards for Privacy of Individually Identifiable Health Information.* Final rule, 45 CFR, pt 160–164. 2002.
17. European Medicines Agency. European Medicines Agency policy on publication of clinical data for medicinal products for human use. EMA/240810/2013. October 2, 2014. http://www.efgcp.eu/downloads/Final%20EMA%20Policy%20Oct%202014.pdf. Accessed September 24, 2015.
18. Malin B. Re–identification of familial database records. *AMIA Annu Symp Proc.* 2006;524–528.
19. Cimino JJ. The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. Appl Clin Inform. 2012;3:392–403.
20. Kaushal R, Hripcsak G, Ascheim DD, *et al.* Changing the research landscape: the New York City Clinical Data Research Network. *J Am Med Inform Assoc.* 2014;21:587–590.

AUTHOR AFFILIATIONS

[1]Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032, USA

[2]Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, New York, NY 10461, USA

[3]Department of Healthcare Policy and Research, Weill Cornell Medical College, New York, NY 10065, USA

[4]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN 37203, USA

[5]Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, TN 37203, USA