Research Article

Cancer Epidemiology, Biomarkers & Prevention

Protecting Privacy in Large Datasets—First We Assess the Risk; Then We Fuzzy the Data

Giske Ursin[1,2,3], Sagar Sen[1,4], Jean-Marie Mottu[5], and Mari Nygård[1]

Abstract

Background: Privacy of information is an increasing concern with the availability of large amounts of data from many individuals. Even when access to data is heavily controlled, and the data shared with researchers contain no personal identifying information, there is a possibility of reidentifying individuals. To avoid reidentification, several anonymization protocols are available. These include categorizing variables into broader categories to ensure more than one individual in each category, such as k-anonymization, as well as protocols aimed at adding noise to the data. However, data custodians rarely assess reidentification risks.

Methods: We assessed the reidentification risk of a large realistic dataset based on screening data from over 5 million records on 0.9 million women in the Norwegian Cervical Cancer Screening Program, before and after we used old and new techniques of adding noise (fuzzification) of the data.

Results: Categorizing date variables (applying k-anonymization) substantially reduced the possibility of reidentification of individuals. Adding a random factor, such as a fuzzy factor used here, makes it even more difficult to reidentify specific individuals.

Conclusions: Our results show that simple techniques can substantially reduce the risk of reidentification.

Impact: Registry owners and large-scale data custodians should consider estimating and if necessary, reducing reidentification risks before sharing large datasets. *Cancer Epidemiol Biomarkers Prev; 26(8); 1219–24. ©2017 AACR.*

Introduction

Much of what we have learned about causes of chronic disease and possible means of prevention is through epidemiologic research on large datasets. Sharing and combining such data is vital for further discoveries of causes of rare diseases. However, it is important that the data are properly deidentified. The access to such data is governed by laws and regulations, namely the European Directive of October 24, 1995, and more recently of April 27, 2016 (1, 2), and Health Insurance Portability and Accountability (HIPAA) regulations 2002 (3–5). With deidentified data, we refer to data where all directly identifying information, such as names, addresses, and social security or personal identification numbers, has been removed, so that the data will appear anonymous to the researcher.

Data that can legally be obtained may contain, in addition to health data, other types of information, such as age, gender, and zip code or other information on city or region. This type of residual information cannot identify individuals on its own. However, such information may be linked to identifiable information in other, often public resources or search interfaces, such as voter registries or social media. A link between these sources can turn residual information into quasi-identifiers, allowing identification of specific individuals. In the United States, voter registration lists used to include voters' names, dates of birth, and zip codes. Linking voter registration lists to residual information in otherwise deidentified sensitive health data might result in reidentifying individuals and violation of privacy. Sweeney (6) combined publicly available data from the Group Insurance Commission in Massachusetts with voter registration lists for Cambridge Massachusetts. Because zip code, age, and gender were in both datasets, Sweeney showed that it was possible to identify the governor of Massachusetts from the insurance data. Others have shown that by knowledge of some additional information, it may be possible to guess the identity of individuals in detailed health data (7). The real number of reidentification attempts with health data may currently be limited, but this remains a concern, in particular with the analysis of vast amounts of data from several sources.

In medicine, much of the recent focus on privacy and big data has been on sharing genetic sequence information (8). However, identifying individuals based on their genetic sequence requires access to a database where these genetic data are already linked to personalized information. This is unlikely. It is usually the nongenetic information in such datasets that poses the threat, because these data can include variables that may be known or can be found publicly.

Increasingly, researchers only obtain access to health data indirectly, that is, the analyses must be done at secure servers behind firewalls, and no data can be removed from the secure server. Methods for such remote access analyses have been developed (9–11). The advantage is that vast amounts of data from different sources can be linked, enabling the researchers to conduct complex analyses on data from many individuals. The disadvantage is that if security to the deidentified data on the secure server were to be breached, the number of quasi-identifiers

[1]Cancer Registry of Norway, Oslo, Norway. [2]Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway. [3]Department of Preventive Medicine, University of Southern California, Los Angeles, California. [4]Simula Research Laboratory, Lysaker, Norway. [5]AtlanModels Team (Inria, IMT-A, LS2N), University of Nantes, Nantes, France.

Corresponding Author: Giske Ursin, Cancer Registry of Norway, P.O. Box 5313 Majorstuen, Oslo 0304, Norway. Phone: 472-245-1300; E-mail: giske.ursin@kreftregisteret.no

doi: 10.1158/1055-9965.EPI-17-0172

is enormous, and reidentification on a massive scale would be possible. Registry owners therefore need to weigh the importance of detailed access to data, even though it will be done securely, to the possible threat to individuals' privacy. Furthermore, for those instances where secure firewalls are not used, and limited data are released to individual researchers, registry owners must in each case assess the possible threat of reidentification.

Reidentification requires that the intruder has access to another dataset that can be used to link quasi-identifiers or identifiers with an individual. Although voter registration lists have been changed, and in the United States, HIPAA regulations now define all dates other than year as protected health information, other such datasets may exist. For example, a bank will have birthdates on all its customers and would know the date when payments were done at the hospital, coinciding with a date when sensitive health information was recorded. The parking company of the hospital might store license plate numbers and dates of parking. Moreover, companies such as Google, Apple, Facebook, and Amazon already have much personal information on individuals and where their smartphones are at any time. Thus, it is not implausible that such datasets exist.

Several methods are available for assessing the risk of reidentification. Estimating this risk means determining the probability that an intruder would discover the correct identity of a single record (12, 13). Once risk of reidentification has been assessed, a number of approaches have been suggested to reduce the possibility of reidentification. An obvious method is to provide less detail in data that are released. Such a method is the k-anonymity protection method (6, 14), where data are aggregated so that there will be k-1 other individuals with the same attributes, that is, within the same equivalence class. Achieving k-anonymization can be done through generalization and/or suppression techniques (14). However, too large aggregation can obscure important details in the data and may render the data useless for analyses. Furthermore, k-anonymity may not always work. Cornell investigators (7) described the problem if all individuals in one of the k-strata or equivalence classes have the same sensitive value on one variable. They also pointed out the challenge if investigators had other background information, and therefore based on such information could recognize specific individuals in a dataset. To overcome these challenges, various additional methods were proposed. L-diversity adds diversity or heterogeneity to the sensitive attributes in each equivalence class with k records, suppressing strata where all individuals have the same value on a sensitive variable. Another method, t-closeness (15), ensures additional diversity (16) by requiring that the distribution of sensitive attributes within each quasi-identifying group should be close to their distribution in the entire database. The value t is the threshold distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the sensitive attribute in the whole database. Other randomization methods (16) involve adding some random factor or noise to the original data.

We took advantage of a large dataset in Norway, where the potentially identifying information was in the date variables, the date of birth, and date of various examinations, information that could potentially be obtained for some women from other sources, including social media. The sensitive data were the outcomes of the various screening examinations. We assessed reidentification risks of the dataset before and after "fuzzifying" the data.

Materials and Methods

The screening programs and threat of reidentification

The Cancer Registry of Norway runs the screening programs on cervical and breast cancer. Both programs started in the 1990s, and the target population for the programs combined exceeds one million every year. The data include dates of when cervical or mammographic exams (screening exams) were obtained, as well as the results from each exam. Some of these women will have developed cancer, and information on the diagnosis is available. Researchers or medical companies may request access to data to assess the effect of various exams and procedures on cancer occurrence or mortality. Requests will include data on dates, as the sequence, type, and outcome of each exam is important. Of the variables available, the largest risk for reidentification of such a dataset is with the date of birth. As for the other dates, the risk may be low with one single exam date, but increases with the combination of specific dates. Sensitive attributes of the dataset are the outcome of some of the laboratory exams and any cancer or precancerous diagnosis.

Anonymization process

Step 1—setting all dates to the 15th of the month. Both the sequence of exams, as well as the time between each exam, are important variables. To preserve this temporal order, we started with a standard registry procedure, which is to remove all days in date variables, and replace them with the number 15. This transformation is a type of k-anonymization, although the value of k is unknown.

Step 2—adding noise or "fuzziness" to dates. The second step in our anonymization process was to add noise to the data. To decide on the magnitude of the noise, we first had to consider the research question being addressed in the study that was going to access the data. The study was examining the association between screening history and subsequent cancer risk. Because screening recommendations are that women get screened every 3 years, we discussed that adding or subtracting some months to every screening date would not alter the validity of the data. However, there was a concern that altering the date with more than a year would not be acceptable.

In our study, a fuzzy factor of 4 was judged as reasonable. This meant that a random integer between −4 and +4 (not 0) was added to the month in every date (months/year) for one person, so that every date was altered by the same random number. The fuzzy factor was different from person to person and was randomly selected. If the process was conducted multiple times, it would give a different random factor for a person each time; it is therefore not possible to reproduce what the process has done. At the same time, to avoid the new dataset being in the same order as the old one, we altered the ID numbers and re-sorted the data. Specifically, the original ID number was replaced by a randomly generated study ID, and the data were sorted on this new variable.

Table 1 shows the data before and after categorizing the date variables, and adding a fuzzy factor. Data at the extreme, for example, very old women, required some additional adjustment, or k-anonymization, to avoid those individuals being identified. We excluded them using a k of 5 when possible. An alternative would have been to assign the value of the closest 5 individuals on that variable, that is, giving all of them the same birth year or age at last exam.

Table 1. Example of data before (a) and after (b) categorizing date variables and adding a fuzzy factor (c)

ID	DOB (day/month/year)	Exam 1 date	Exam 2 date	Diagnosis date
A. Entries in D1 realistic dataset				
01071972 23456	1/7/1972	2/8/2000	10/11/2004	21/1/2007
31051970 65432	31/5/1970	5/8/2005	1/12/2008	21/2/2011
03041960 45678	3/4/1960	5/1/1995	10/2/1998	—
20021981 87654	20/2/1980	20/9/2006	10/8/2009	23/1/2010
B. Entries in D2 after categorizing date variables (applying k-anonymization, setting every date to the 15th)				
001	15/7/1972	15/8/2000	15/11/2004	15/1/2007
002	15/5/1970	15/8/2005	15/12/2008	15/2/2011
003	15/4/1960	15/1/1995	15/2/1998	—
004	15/2/1980	15/9/2006	15/8/2009	15/1/2010
C. Entries in D3 anonymized and "fuzzified" (the fuzzy factors assigned in this case were +3 for ID001, −3 for ID002, +4 for ID003, and −1 for ID004)				
1023	15/10/1972	15/11/2000	15/2/2005	15/4/2007
9875	15/2/1970	15/5/2005	15/9/2008	15/11/2010
4567	15/8/1960	15/5/1995	15/6/1998	—
2345	15/1/1980	15/8/2006	15/7/2009	15/12/2009

Estimating reidentification risk in original, k-anonymized and fuzzified data

The fundamental approach to understanding the risk of reidentification is by measuring the probability that an intruder would discover the correct identity of a single record (12, 13). For example, if an intruder knows the birth date May 5, 1972, and the exam date April 19, 2005, of Anne, then he/she needs to find/match unique records in a database with the combination (1). Finding an exact match for such a combination would reveal the values of the sensitive attributes such as the diagnosis from the exam. If there are f matching records for a combination of variables, this gives a risk of reidentification of Anne's record $= 1/f$.

However, the variables for birth date and exam dates for each individual could be recoded to a range of years. In this case, Anne's dates would fall into the range of years 1972 to 1975 and 2003 to 2006. Across the database, these date ranges represent equivalence classes. The probability of reidentifying a patient in the smallest equivalence class in a database represents the overall risk of the database. There can be $i = 1 \ldots I$ equivalence classes in a database, where I is the total number of equivalence classes in the database. If all values of a variable are unique, then I is equal to the total number of records making every person uniquely identifiable. However, this is unlikely as some people share values of attributes such as birth dates or exam dates. The number of records in an equivalence class is denoted by f_i. For an attacker who wants to identify a single person given some background knowledge about a quasi-identifying variable, the overall probability of reidentification is computed as the minimum value of $1/f_i$ across all equivalence classes $= 1/(min(f_i)$. For very large databases, determining the uniqueness of values for different attributes is computationally expensive; hence, tools to determine reidentification risk often use uniqueness estimates (17).

Our approach to deidentification and fuzzifying did not replace dates with ranges, but changed all days to 15th of a month (step 1), and added a fuzzy factor to month (step 2). Therefore, the size of the equivalence classes after step 1 is determined by how many people were born the same month, had a diagnosis, or were censored in the same month.

We evaluated the risk of reidentification of a dataset using the tool ARX (18). This tool can be downloaded to a desktop and can handle datasets of several million records to evaluate risks. The three principal steps to evaluate reidentification risk using ARX are presented below:

Step 1. Import data into ARX: The dataset is imported into ARX and saved.
Step 2. Qualifying attributes to be identifying, quasi-identifying, or sensitive: The user classifies attributes as identifying, quasi-identifying, or sensitive. For instance, a national ID or social security number is identifying, birth date is quasi-identifying, and diagnosis is sensitive.
Step 3. Analysis of reidentification risk: ARX computes reidentification risk based on three different attacker models: (i) the prosecutor scenario; (ii) the journalist scenario; and (iii) the marketer scenario. In the prosecutor model, it is assumed that the attacker knows that data about the targeted individual is contained in the database. In the journalist model, such background knowledge is not assumed. In the marketer model, it is assumed that the attacker is not interested in reidentifying a specific individual but that he/she aims at attacking a larger number of individuals. An attack is only considered successful if a large portion of the records could be reidentified. The risk analysis in ARX uses estimates of uniqueness either based on the sample (17) or on a superpopulation model (19). ARX determines the lowest, maximum, and average risk of reidentification for the prosecutor model and presents estimated risk for the journalist and marketer models. In this article, we use the prosecutor scenario, that is, the most aggressive scenario, where the goal is to reidentify specific individuals. We therefore present only the ARX estimate of the prosecutor risk.

The data used in this article consisted of a random sample of women who attended cervical cancer screening and diagnostic work-out exams in Norway from 1992, when the screening program started, until 2014. For each individual, the following variables were available: date of birth, date of emigration or death, and the following information on each cervical examination: (i) date of exam; (ii) type [cytology/histology/human papillomavirus (HPV) exam]; (iii) diagnosis/outcome (cytology, histology or HPV result); (iv) laboratory responsible for the diagnosis (code from 1–21), geographic health care region in Norway of the laboratory (South-East, West, Middle, North). Cervical cancer was additionally described by stage at diagnosis (I–IV, with

various substages a1–b2). The censoring date was the date of emigration, death, or December 2014 (whichever came first).

Ethical approval

This research project was approved by the South East Regional Committee for Medical and Health Research Ethics. The data provided for this project had already been k-anonymized and fuzzified once prior to release. To conduct the project, we therefore started by assigning a fabricated day (random number between 1 and 28–31 depending on the month) to each date to obtain a realistic dataset.

Results

The dataset contained 5,693,582 records of screening related examinations taken by 911,510 distinct women. The birth dates of the women ranged from March 1905 to February 1996.

The risk of reidentification was assessed for the following datasets:

D1. Realistic dataset of women attending cervical cancer screening in Norway.

D2. k-Anonymization of the dataset D1 by changing all dates in the dataset to 15th of the month.

D3. Fuzzifying the month in D2 by adding a random factor between −4 and +4 months to each month as described above.

The quasi-identifying attributes that a prosecutor is most likely to use in this dataset are the birth date, exam dates, and the censor date.

Reidentification risk analysis of D1

In the prosecutor scenario, it is assumed that the attacker already knows that data about the targeted individual are in the dataset and would know an individual's birth, exam dates, and diagnosis/censor date. The realistic dataset D1 had a very high average prosecutor risk of 97.1% to identify a person (Fig. 1A), as the combination of all three "exact" dates uniquely identified many women. Over 94% of records were affected by the highest risk of 100%. However, if a prosecutor only had knowledge of birth date or knowledge of birth date and censor date, the average prosecutor risk was lower (0.5% and 2.5% respectively, not shown in figure).

Reidentification risk analysis of D2

There is a significant impact on prosecutor risk when D1 is transformed to D2 where all days in a date are transformed to the 15th of the month. The average prosecutor risk drops to 9.7% (as shown in Fig. 1B), as many individuals in D2 will share the same date for birth, diagnosis, and censoring. However, 6% of all records still had the highest risk of 100%. This was because of outlier patients that have more exams than the majority of the population. If a prosecutor knew the number of exams of a specific patient, this could be a way of identifying a person. One way to suppress this could be to exclude patients who have a number of exams greater than a certain threshold. If birth and censor dates were the only quasi-identifiers, then the prosecutor risk drops to 1.48%. This drops further to 0.0193% if only birth date was known.

Reidentification risk analysis of D3

Adding a random fuzzy factor of −4 to +4 months to the month of date in D2 yielded similar average prosecutor risks from ARX as in D2. We obtained an average prosecutor risk of 9.8% as shown in Fig. 1C. Again, 6% of all records are affected by the highest risk of 100% just like in D2. Similarly, using birth date and censor date as quasi-identifiers gave a prosecutor risk of 1.48% for D3 and 0.0194% if only birth date was a quasi-identifier. In other words, there were as many unique records in D3 as in D2. However, a prosecutor would have a more difficult time linking those records from D3 back to any individuals, as all the months had been altered by a random fuzzy factor.

Discussion

This study shows that risk of reidentification can be substantially reduced by simple procedures. In our example, replacing the day in all date variables with a fixed number, such as 15, greatly reduced the reidentification risk. However, there were still a number of individuals with a large risk of reidentification. We were concerned that doing additional k-anonymization, that is, further collapsing categories or dates to years would compromise the validity of the data by obscuring details useful for medical studies. We therefore added a fixed fuzzy factor to all the months in the date variables. In the calculation of reidentification risk, this did not alter the average prosecutor risk, because this did not result in fewer unique records. However, the possibility of reidentifying any specific individual would presumably be even lower with this very simple addition.

Any modification of data by k-anonymization or adding noise to the data is a potential threat to data validity. Development of research question–specific protocols should therefore be done in collaboration with researchers analyzing the data to determine what type of anonymization procedure is possible.

It is standard procedure in registries and for owners of large datasets to obscure dates by altering all days in a date to 15. This was the first step in our procedure. This type of categorization is a type of k-anonymization, but as we showed, there were still a number of unique records in the data, with a high risk of being reidentified. This was not surprising and confirms that dates that include months are indeed identifying for a number of individuals, as HIPAA regulations indicate. However, to analyze these screening data, the sequence of events was necessary, and further generalization or categorization on dates could threaten the validity of the data. Another possibility would be to remove all calendar years and provide all data simply as days from birth. This could be useful in a survival analysis of a cohort with a relatively small window of enrollment. However, in our instance, because screening methods have changed over time, this was not an acceptable solution, as researchers needed to examine changes over time, so an approximate estimate of when the exams took place was important. In many scenarios, it is necessary that the interval of number of days between events is preserved for studies, such as survival analysis. In a study on deidentifying nursing notes, the authors describe software that reads free-text medical records and automatically shifts dates, including date of birth, by a patient-specific random number of days, while preserving the day of week and season (20). Our method is similar, but used on registry data, not individual medical records.

In our example, the dataset contained relatively limited amount of data items per individual. If the screening data had been combined with other variables, a larger fuzzy factor or a

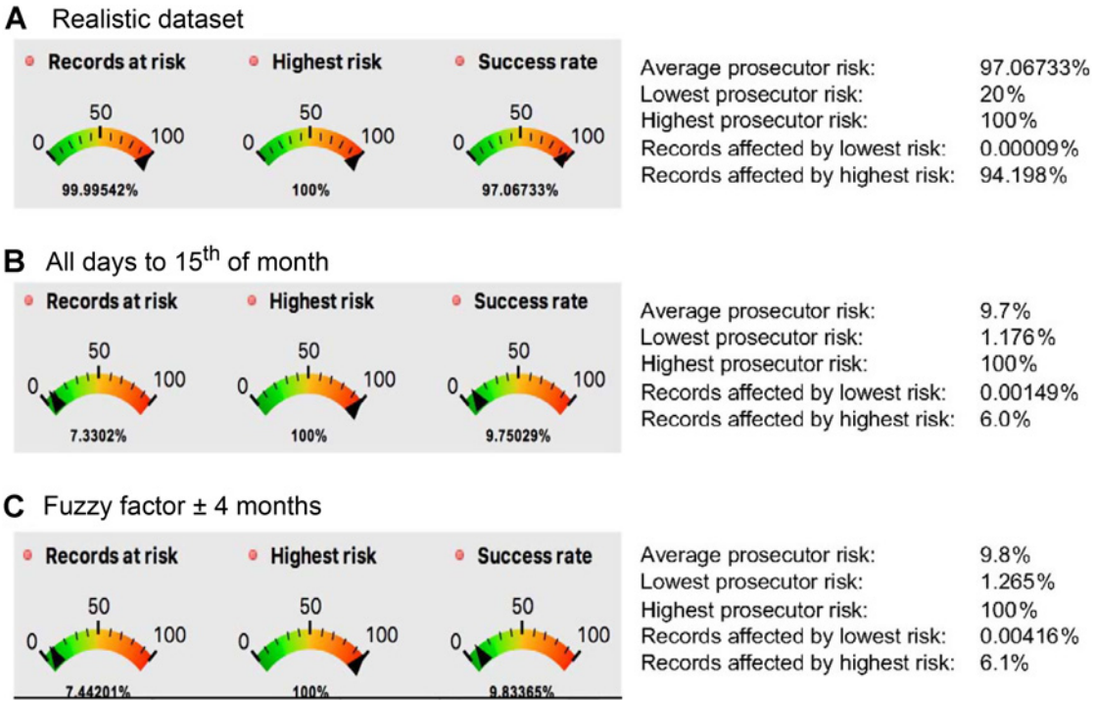

Figure 1.

ARX reidentification risk analysis. **A,** Realistic dataset. **B,** All days in a date changed to 15th. **C,** All months perturbed by a random number between −4 and +4. Years updated to ±1 accordingly.

fuzzy factor combined with additional types of anonymization techniques could have been necessary to reduce reidentification risks. Protocols and tools for such additional anonymization are readily available (18).

Our data contained only women. However, if this had been a study of colorectal cancer screening, it would have been important to assess reidentification risks separately for men and women. Although gender may seem as an innocuous variable by itself, combined with other variables, it could have a large impact on reidentification risk.

Another advantage of adding a random fuzzy factor to each date was that the data have actually been altered, and it is truly not the person's own data that are being released.

A challenge with this method could be if the fuzzy factor was set to become large, say −12 to +12. Occurrences in the data at one time point could then result in appearing one year earlier in one person, and one year later in another person. In our case, the −4 to +4 range seemed reasonable. Another challenge could be that adding fuzziness could obscure important interrelationships within the data. If a goal is to study the associations between subjects, for example, between individuals who share a sexual partner, then this method should be modified, or the range set very narrow. Finally, we assumed the prosecutor had access to only one exam date. If the prosecutor had access to multiple exam dates in addition to birth date and diagnosis/censoring date, then the reidentification risks may be underestimated.

There is great concern among statisticians and epidemiologists that adding noise to the data will obliterate the signal. However, a review of the cancer epidemiology literature will show that in most instances, epidemiologists will classify both exposure, confounding and outcome variables into more or less broad categorical variables. Although this can be a result of detailed analysis of each variable, often it is not. Thus, there should be substantial room for discussion on what detail is sufficient for each research question, and what level of noise would be acceptable to not compromise the value of the data.

Large datasets are increasingly being analyzed at secure servers, behind firewalls. This is becoming the new standard for combining large data. Such systems often make it impossible for researchers to pull out anything except tabular results. However, in some systems, it may still be possible to look up individual records. Thus, a researcher so inclined could obtain information on specific individuals simply by combining the data in a similar fashion to Sweeney (6). In addition, there has been a number of recent hacker attacks on presumed secure servers. If the server was breached, it would be very important that the data are not easily reidentified.

We suggest that data custodians consider running datasets through one of the available tools to assess reidentification threats prior to release of data. In addition to ARX, there are large number of existing tools, such as the commercial PARAT (21), UTD Anonymization Toolbox1 (22), Cornell Anonymization Toolkit (23), sdcMirco23 (package for the *R* statistics software; ref. 24), and μ-Argus24 (25).

Large data analyses are important, and there is a great need to share large datasets; however, this needs to be done safely, and such that the threat of reidentification is minimal. We have shown

an example of applying simple anonymization methods to a large screening dataset. Tailoring existing anonymization methods to the data and the research protocols may be the best method for ensuring privacy of the individuals who have provided the data.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: G. Ursin, S. Sen, M. Nygård
Development of methodology: G. Ursin, S. Sen, M. Nygård
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Nygård
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): G. Ursin, S. Sen, J.-M. Mottu, M. Nygård
Writing, review, and/or revision of the manuscript: G. Ursin, S. Sen, J.-M. Mottu, M. Nygård
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): G. Ursin, S. Sen, M. Nygård
Study supervision: G. Ursin, M. Nygård

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 22, 2017; revised April 7, 2017; accepted April 21, 2017; published OnlineFirst July 28, 2017.

References

1. Bossi J. European directive of October 24, 1995 and protection of medical data: the consequences of the French Law Governing Data Processing and Freedoms. Eur J Health Law 2002;9:201–6.
2. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
3. U.S. Department of Health and Human Services. The HIPAA Privacy Rule. Washington, DC: U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/.
4. Annas GJ. HIPAA regulations - a new era of medical-record privacy? N Engl J Med 2003;348:1486–90.
5. Schoppmann MJ, Sanders DL. HIPAA regulations. J Am Coll Radiol 2004;1:728–33.
6. Sweeney L. k-Anonymity: a model for protecting privacy. Int J Uncertain Fuzz 2002;10:557–70.
7. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data 2007;1:3.
8. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014;15:409–21.
9. Sparks R, Carter C, Donnelly JB, O'Keefe CM, Duncan J, Keighley T, et al. Remote access methods for exploratory data analysis and statistical modelling: privacy-preserving analytics. Comput Methods Programs Biomed 2008;91:208–22.
10. Thompson G, Broadfoot S, Elazar D. Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics. Geneva, Switzerland: United Nations Economic Commission for Europe; 2013. Available from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf.
11. Heldal J, Monstad E, Risberg T, Risnes Ø. The RAIRD Project: Remote Access Insfrastructure for Register Data.Geneva, Switzerland: United Nations Economic Commission for Europe; 2015.
12. Emam KE, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the risk of re-identification of patients from hospital prescription records. Can J Hosp Pharm 2009;62:307–19.
13. Dankar FK, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak 2012;12:66.
14. Samarati P. Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 2001;13:1010–27.
15. Li N, Li T, Venkatasubramanian S, editors. t-Closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering; 2007 Apr 15–20; Istanbul, Turkey. New York, NY: IEEE; 2007.
16. Aggarwal CC, Yu PS. Privacy-preserving data mining: models and algorithms. Springer US; 2008. New York, NY.
17. Haas PJ, Naughton JF, Seshadri S, Stokes L. Sampling-based estimation of the number of distinct values of an attribute. In: Proceedings of the 21th International Conference on Very Large Data Bases; 1995 Sep 11–15; San Francisco, CA. Burlington, MA: Morgan Kaufmann Publishers Inc.; 1995. 311–22.
18. Prasser F, Kohlmayer F, Lautenschlager R, Kuhn KA. ARX–A comprehensive tool for anonymizing biomedical data. AMIA Annu Symp Proc 2014; 2014:984–93.
19. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: the ARX data anonymization tool. In: Gkoulalas-Divanis A, Loukides G, editors. Medical Data Privacy Handbook. Cham, Switzerland: Springer International Publishing; 2015. p.111–48.
20. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
21. Privacy Analytics. PARAT De-Identification Software. Available from: https://www.privacy-analytics.com/software/privacy-analytics-eclipse/.
22. UT Dallas. UT Dallas Anonymization Toolbox. Available from: http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php.
23. Xiao X, Wang G, Gehrke J. Interactive anonymization of sensitive data. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. 2009 Jun 29–Jul 2; Providence, RI. New York, NY: ACM; 2009. p.1051–4.
24. Templ M, Kowarik A, Meindl B. Statistical disclosure control for micro-data using the R Package sdcMicro. J Stat Softw 2015;67:1–36.
25. μ-Project. μ-ARGUS manual. Available from: http://neon.vb.cbs.nl/casc/mu.htm.