Published in final edited form as:

Med Care. 2013 August ; 51(8 0 3): S58–S65. doi:10.1097/MLR.0b013e31829b1d10.

Privacy Technology to Support Data Sharing for Comparative Effectiveness Research: A SYSTEMATIC REVIEW

Xiaoqian Jiang[*,+], **Anand D. Sarwate**[**,+], and **Lucila Ohno-Machado**[*]

Xiaoqian Jiang: x1jiang@ucsd.edu; Anand D. Sarwate: asarwate@ttic.edu; Lucila Ohno-Machado: lohnomachado@ucsd.edu

[*]Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA 92093

[**]Toyota Technological Institute at Chicago, Chicago, IL 60637

Abstract

Objective—Effective data sharing is critical for comparative effectiveness research (CER), but there are significant concerns about inappropriate disclosure of patient data. These concerns have spurred the development of new technologies for privacy preserving data sharing and data mining. Our goal is to review existing and emerging techniques that may be appropriate for data sharing related to CER.

Material and methods—We adapted a systematic review methodology to comprehensively search the research literature. We searched 7 databases and applied three stages of filtering based on titles, abstracts, and full text to identify those works most relevant to CER.

Results—Based on agreement and using the arbitrage of a third party expert, we selected 97 articles for meta-analysis. Our findings are organized along major types of data sharing in CER applications (i.e., institution-to-institution, institution-hosted, and public release). We made recommendations based on specific scenarios.

Limitation—We limited the scope of our study to methods that demonstrated practical impact, eliminating many theoretical studies of privacy that have been surveyed elsewhere. We further limited our study to data sharing for data tables, rather than complex genomic, set-valued, time series, text, image, or network data.

Conclusion—State-of-the-art privacy preserving technologies can guide the development of practical tools that will scale up the CER studies of the future. However, many challenges remain in this fast moving field in terms of practical evaluations as well as applications to a wider range of data types.

1 Introduction

The purpose of Comparative Effectiveness Research (CER) is to inform patients, providers, and decision-makers about the effectiveness of different interventions[1]. CER promises enormous societal benefits by promoting new scientific evidence in medicine, speeding up clinical discoveries, and enabling cost- and time-effective patient care. To achieve these goals, CER researchers must obtain access to a wide range of information (e.g., demographics, lab tests, genomic data, and outcomes) from a variety of population groups. Institutions face fundamental challenges in how to share data with researchers or with the public; they must balance the privacy of patient data with the benefits of CER.

[+]Xiaoqian Jiang and Anand Sarwate contributed equally to this paper

The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) sets standards for the privacy and security of health records in United States[2]. HIPAA defines two approaches for de-identification: *Expert determination* and *Safe Harbor*. The *Expert Determination* approach requires that a statistician certify that the re-identification risk in the data is “sufficiently low”. The *Safe Harbor* approach, on the other hand, explicitly requires the removal and suppression of a list of attributes[3]. The Department of Health and Human Service (HHS) has recently issued revised guidance on methods for de-identifying protected health information (PHI) [*]. These changes clarify the de-identification standard and how to perform de-identification, but do not change the existing standards. For the expert determination approach, the new guidance defines key concepts such as covered entities, business associates, and acceptable risk, explains standards for satisfying the standard, and gives examples of how expert determination has been applied outside the healthcare context, such as within government statistical agencies like the Bureau of the Census. For the safe harbor approach, the guidance provides more examples, including when zip codes and elements of date can be preserved in the de-identified data, as well as how to use data use agreements when sharing de-identified data. These statutory requirements are included in the Appendix.”

There are numerous controversies on both sides of the privacy debate regarding these HIPAA privacy rules[4]. Some people believe that protections in the de-identified data under HIPAA are not sufficient[5] – a 2005 national consumer health privacy survey also showed that 67% of national respondents remain concerned about the privacy of their personal health information[6], indicating a lack of public trust in the protection offered by HIPAA de-identified data[7]. On the other side, others contend too many privacy safeguards hamper biomedical research, and implementing these safeguards precludes meaningful studies of medical data that depend on suppressed attributes (e.g., epidemiology studies in low-population areas or geriatric studies requiring detailed ages over 89[3]). They also worried about the harm caused by privacy rules – they could erode the efficiencies offered by computerized health records and possibly interfere with law enforcement[4].

In practice, privacy always comes with a loss of *utility* – perfect privacy is only possible when no data are shared. However, this measure of utility is application-dependent. In this paper, we focus on data sharing problems that may arise in CER applications. For such applications, we can measure utility by metrics such as classification accuracy and/or calibration. Some privacy-preserving operations may destroy too much information to achieve these target goals. *Privacy metrics* allow institutions to evaluate the tradeoff between the improvements from integrating additional data and the privacy guarantees from the privacy-preserving operations.

In order to realize the benefits of improved care via CER, data holders must share data in a way that is sensitive to the privacy concerns of patients. We synthesize and categorize the state of the art in privacy-preserving data sharing, a topic that has sparked much research in the last decade. We expect this study can guide CER researchers in choosing a privacy method, inform institutions in developing data sharing agreements, and suggest new directions for privacy researchers.

2 Material and methods

2.1 Search Strategy

We adapted a systematic review methodology, suggested by Centre for Reviews & Dissemination guide[8], to review the research literature. Figure 1 illustrates our flow of

[*] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

information through the different phases of this review process. We chose to query seven databases, shown in Table 1. We used the basic format of posing broad queries to capture as many relevant articles as possible and then applied targeted inclusion criteria to find those works relevant to CER. The search included documents published by 2/1/2012. The online Supplemental digital content of this manuscript provides details of our methods. Because many relevant studies were identified in the computer science (CS) literature, we provide some explanation for CS-specific terminology in Table 2 for the benefit of readers.

2.2 Synthesis

The details of our study are available in Supplemental digital content. To contextualize our findings for privacy-preserving data techniques, we outline three examples of typical situations that may arise in the context of CER.

Institution-to-institution—Researchers from Institution A want to study the benefits of minimally invasive surgery of their own patients and patients at Institution B, another hospital that routinely use Da Vinci Robotic Surgical system to conduct minimally invasive surgery for cardiac patients. To provide information about their patients, Institution B generates an anonymized data table, together with a data-use agreement limiting access to authorized researchers at Institution A.

Institution-hosted—Institution A wants to make collected data about diabetes care available to researchers (internal or external), who study diabetes complications in stoke. Instead of sharing data directly with individual researchers, Institution A sets up a hosted data warehouse to answer the queries of researchers through a secure web interface (e.g., clinical data warehouse).

Public release—Institution A wants to make collected readmission rates of cardiac patients (within 30 days of discharge) publically available for the purpose of safety surveillance. Statisticians at the Institution A analyze the raw data and generate a number of statistical analyses, summaries, and tables derived from the data to be published.

In the CER context, the above-mentioned examples represent different modalities for data sharing. When data are shared directly between institutions, they are covered by a data use agreement. In this scenario, the major challenge is protecting the data confidentiality during the transfer process. Regarding the clinical data warehouse scenario, data stewards implement a controlled interface to the sensitive data so that the answers to the queries are protected (similar to those existing ones like i2b2[9] and CRIQueT[10]). For the public dissemination case in our last example, the type of data that can be released is much more limited than in controlled settings (e.g., no individual patient records). It is important to choose appropriate anonymization models, techniques, and algorithm parameters in conjunction with data use agreements to avoid information breaches.

2.3 General metrics and methods

The word “privacy” has different meanings depending on the context. What we refer to as “privacy” in this paper often goes by “confidentiality” in the statistical literature[11,12]. The goal of privacy-preserving data sharing is to manipulate the original data in such a way as to prevent re-identification of identities or sensitive attributes. There are many methods for publishing versions of the original data, such as suppression of unique elements, top/bottom coding to limit ranges of values, generalization by merging categories, rounding values to limit uniqueness, and adding noise. Another approach is to simply release synthetic data that in some way “looks like” the real data – methods for this include sampling and partial data substitution. If releasing the original data is not necessary or is considered too risky, summary statistics or sub-tables of the data can be generated from the original data. More complex anonymization and sanitization algorithms are often built on these basic operations using structural properties of the data sets. Fayyoumi et al. reviewed various techniques on statistical disclosure control and micro-aggregation techniques for secure statistical databases[13].

Altering the original data makes the disclosed data less useful, so a key element of privacy technologies is providing a metric for the level of protection[14–16]. This allows empirical evaluations of the difference in utility between the original and manipulated data. There are many surveys of privacy operations and metrics[12,17–20], but they do not address applications in CER.

The choice of a privacy model or technology depends on the perceived threats to confidentiality; it is therefore important to specify to whom the data are being shared and what sort of external restrictions are placed on the recipients of the data. Many proposed methods for privacy-preserving data analysis or sharing do not provide any formal or quantifiable guarantees of privacy; instead, they claim that because the shared data are sufficiently “different” from the original data, they are inherently private. A useful privacy-preserving data sharing method should specify the threats as well as quantify the level of protection provided. Quantification of the privacy risk is important because it allows the system designer to compare different algorithms and evaluate the tradeoffs between privacy and the utility of sanitized data[19].

We can divide the privacy metrics proposed in the literature into two categories: syntactic and semantic. Syntactic metrics are defined in terms of properties of the postprocessed “sanitized” data. For example, *k-anonymity*[21] guarantees that, for any combination of feature values, if there is one person with those features, there are at least *k* with the same feature values. To achieve this goal, original feature values may be merged (e.g., lab tests are reported as ranges rather than values). The anonymization system Datafly[22] uses *k-anonymity*, and many government agencies use a “rule of k” (another version of k-anonymity) to determine if data are anonymized. Other metrics such as *l*-diversity[23] and *t*-closeness[24], or *m*-invariance[25] provide related guarantees on the level of masking. There is extensive literature about attacks on these privacy models[26–31].

Semantic privacy measures are defined in terms of the properties of the process of data sanitization. The most studied version of semantic privacy is *differential privacy*[23], which provides a statistical guarantee on the uncertainty in inferring specific values in the data. In syntactic privacy, the released data set satisfies particular privacy conditions, whereas in semantic privacy, the process guarantees privacy, regardless of the underlying data. However, differential privacy is still subject to inferential attacks[32]. Another model for privacy risks is ∂-presence[33,34], which models the effect of public data on inferring the presence of individuals in a data set.

Regarding assumptions on threats, syntactic privacy methods either assume that the recipient of the data knows nothing about the individuals in the data or assume that the adversaries have limited knowledge. The former is a dangerous assumption, especially for public release data sets, since there are many publically available datasets that can be used to launch a so-called *linkage attack* (see Table 2). The second approach is difficult because it requires modeling the knowledge of an unknown adversary. By contrast, *differential privacy* guarantees that an adversary with full knowledge of all but one individual’s data will still have difficulty inferring the data of that individual. While it is a robust definition in this sense, many differentially private algorithms are not practical for use on small or moderate-sized data sets[35].

These models of adversarial knowledge are pessimistic in that they assume the recipient of the data intends to re-identify individuals. While this may be a reasonable assumption for some users of public-use data sets such as Medicare billing data, in other scenarios the data holder can issue enforceable data use policies that can hinder re-identification attempts. Prohibiting re-identification research is a mistake that can prove very costly in the near future, but developing accompanying policies limiting access to sensitive data via data sharing agreements or hosted enclaves can reduce the chance of inadvertent identity disclosure.

3 Results

3.1 IDENTIFIED PRIVACY METHODS FOR CER APPLICATIONS

In institution-to-institution sharing, the privacy risks are not as uncontrolled as they are in public data release. We found several articles in the literature that suggest algorithms to address this kind of data sharing scenario. The kind of protections provided and the resulting utility of the data are different for these methods. There is an extensive literature on *k-anonymizing*[21] a data set prior to publication or sharing[36–44], and there are also implementations satisfying other syntactic privacy measures[43, 44, 73–79]. A k-anonymization approach was proposed by El-Emam et al[45] in the context of (public) data publishing for medical data. Another recent promising approach for linking data sources in an federated system was proposed by Mohammed et al[46]. The effect of these approaches on utility can vary. Some work has focused on enhancing utility through post-processing[47] or evaluating the effect of anonymization on specific statistical tasks[48].

Perturbing the data table prior to information exchange can also protect privacy. The perturbation can be chosen to provide a privacy guarantee or to maintain a certain level of utility. Some of this work arose from the statistical literature and used statistical measures for measuring privacy, such as posterior odds[49] or other metrics[50]. In contrast with the syntactic investigations of k-anonymity, noise addition has a more directly measurable impact on utility, and several studies investigated the effect of the noise on the utility of the data. *Differential privacy* has been proposed for sharing anonymized tables of data such as contingency tables[51]. More advanced methods with utility analyses for differential privacy have been developed using wavelet transforms[52].

Secure Multiparty Computation (SMC) allows multiple parties to perform computation on their private data to evaluate some function of their common interest[53, 54]. Basically, these approaches apply a set of cryptography motivated techniques to ensure that data sources collaborate to obtain results without revealing anything except those results[55]. SMC techniques have been developed for classification[56,57], clustering[58], association rule mining[59], and data disclosing for disease surveillance[60], which demonstrated powerful privacy protections. A detailed classification of these algorithms was provided by Xu[61]. A recent paper[62] suggested privacy and collaborative data mining (i.e., CER data mining) can be achieved at the same time when the computational task is well-defined.

In an institution-hosted framework, CER researchers have access to the data through an interactive mechanism that can monitor and track their privacy usage. This is a preferable arrangement when the information that needs to be shared is not known in advance or may change over time. While queries can be processed on a special anonymized data set created using the techniques mentioned in the previous section[63], there are some approaches to explicitly handle interactive queries. Syntactic privacy methods generally do not address interactive methods, although recent work has reported on a framework for instant anonymization[64].

Differential privacy was first proposed in the context of interactive queries[65–68]. Typically, privacy is enforced through returning noisy responses to queries, although theoretical work has proposed more complex query processing[69]. This privacy model has been incorporated into query languages for data access[70,71] and MapReduce, which is the system used by Google and others to perform computations on large data sets[72]. In the medical informatics community, noise addition has been proposed for exploratory analysis in a clinical data warehouse[9] and differential privacy has been proposed for count queries[50]. Other approaches to online analytical processing (OLAP) use statistical measures of privacy[73].

To prepare data for public release, the data custodians need to set the confidential level high enough to protect sensitive patient privacy from breaches because a broad disclosure of health data poses a much more significant privacy breach risk than previous scenarios of institution-to-institution and institution-hosted data access. Recent examples stemming from data shared by Netflix and AOL showed that simply removing identifiers or naïve aggregation may not be enough, and that more advanced de-identifying techniques are needed. In the Netflix case, individuals in an anonymized publicly available database of customer movie recommendations from Netflix were re-identified by linking their ratings with ratings in the Internet movie rating web site IMDB[74]. In the AOL case[75], a reporter re-identified an AOL user in released “de-identified” search queries, and revealed that a combination of several queries was enough to narrow the searcher’s identity to one particular person[76]. Privacy breaches are often reported in the popular press [77], and represent a strong disincentive for sharing data.

To avoid privacy pitfalls and to mitigate risk, numerous articles have been published to setup a foundation of privacy preserving data publishing for general and specific applications[78–80]. One line of approach, including k-anonymity, as introduced earlier, manipulates the data to merge unique individuals, sanitizing tables through table “anonymization”[33,81,82] (i.e., generalization or suppression) before publication. Another approach to this kind of data sharing is producing synthetic data, which are supposed to capture the features of the original data. In this context, this would involve generating fictitious patients, who “look like” the real patients. Several methods have been proposed that do not explicitly quantify privacy[83], adopt novel risk measures[84], or use a blend of anonymized and synthetic data[85]. Others create compact synopses, including wavelets[52], trees[86], contingency tables[87,88], and compressed bases[89], and sample synthetic data from the synopsis. In the literature on *differential privacy*, synthetic data generation has attracted significant interest for a theoretical standpoint[90] (see also follow up work[91]), but there are limited studies to evaluate the usefulness of *differentially private synthetic data* in real world applications[92,93].

3.2 RECOMMENDATIONS

Although we identified a few privacy technologies that can facilitate CER research, in order to realize the full potential of CER studies in a privacy-sensitive way, more work has to be done to bridge the gap between CER researchers, statisticians, informaticians, and computer scientists. In particular, these communities can work together to develop more precise formulations of CER data sharing problems, benchmarks for privacy and utility, and realistic expectations of how much protection must come from technology (algorithms) versus policy (use agreements).

CER researchers can contribute by more concretely specifying their data sharing needs. For example, for a large multi-site study, what information really needs to be shared? Perhaps a preliminary assessment would show that some portions of the raw records are not needed. By developing canonical data sharing and study examples, designers can develop algorithms that are tuned to those settings.

Statisticians who work on CER studies are best positioned to specify the kinds of inference procedures they need to run on the data. This in turn will inform algorithm design to help minimize the distortion in those inferences while still preserving privacy. Not enough work has been done to develop meaningful utility metrics. There is a rich literature on enhancing data utility during anonymization[47,94–97], however, the metrics vary widely[24,97,98]. It is important to develop standards for utility and data quality that are relevant for CER applications. These in turn can dictate the kinds of policy protections and algorithmic parameters to use in anonymization. By integrating the statistical task to be performed with the data sharing structure for the CER study, researchers can develop a concrete and well-specified problem for algorithm designers.

The last piece is to develop a set of comprehensive benchmarks on standardized data that other research communities such as the machine learning and computer vision communities use to compare and validate novel models. Such benchmarks can be used to provide head-to-head comparisons of existing privacy-preserving technologies. This requires the work of all parties to find concrete examples and corresponding data for each of these canonical data sharing examples. This research reproducibility will steer the development of algorithms by making it clear which ones are successful.

The field of CER evolves rapidly. New emerging applications may involve new data types and there might be no privacy standards to protect them. Such a gap between policy and technology calls for substantial future development of new standards of healthcare data privacy protection for genomic data[99–103], set-valued data[104], time series data[105], text data[106,107], and image data[108], which have not been adequately studied in the privacy perspective.

4 Discussion

As we described in the previous section, many of the new anonymization and privacy-preserving data publishing techniques can be applied to scenarios of interest in CER. Some of these approaches are still under active development, and choosing privacy metrics and algorithms will depend not only on the data sharing structure but also on the specific data to be shared and policy considerations. Data sharing agreements can mitigate the loss of utility in anonymized data at the expense of more policy oversight. Entities such as an Institutional Review Boards (IRB) exist in many organizations and can provide guidelines on data use to prevent researchers from inappropriately using the shared data to re-identify individuals. For example, *in institution-to-institution data sharing arrangements*, enforceable contracts can be signed between the institutions to guarantee oversight of the shared data as well as to describe appropriate uses for the data. For hosted-access models, users who wish to access the data could sign use agreements that restrict how they can disclose the information; such models are used routinely by government agencies in data enclaves such as the National Opinion Research Center[109]. The greatest danger comes from public dissemination of data, where there can be no reasonable restrictions placed on the public’s use of the data. In such a setting, privacy protections must be correspondingly stronger and more comprehensive.

Ultimately, the choice of privacy level will be dictated by a combination of policy considerations applied to these tradeoffs. Improved data governance policies and data sharing agreements could help mitigate the impact that privacy-preserving operations have on utility by providing a technological and legal framework for preventing misuse of patient data. Privacy-preserving data manipulation is an important part of a larger data-governance ecosystem that encompasses informed consent, data use agreements, and secure data repositories.

While there is a substantial and growing literature on privacy preserving techniques in computer science, statistics, social science, and medicine, many of these works are not directly applicable to the CER context. We surveyed state-of-the-art literature to find relevant papers, sort them, and make recommendations based on three major axes of CER applications (i.e., institution-to-institution, institution-hosted, and public release). Despite encouraging findings, we also identified a serious gap between theory and practice. To close this gap, CER researchers should specify statistical objectives from data sharing and privacy researchers should develop methods adapted to these objectives. New methods will be needed to handle more complex forms of data that arise in healthcare.

Obtaining real clinical benchmark data and initiating competitions between privacy technologies using that data, researchers can help build a healthy ecosystem between the CER and privacy communities. Such an exchange can encourage the sharing of ideas and development of real testable standards and benchmarks. Addressing these issues and overcoming challenges will catalyze the CER studies of the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Federal Coordinating Council for Comparative Effectiveness Research. [Accessed August 30, 2011] http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf
2. [Accessed Feburary 20, 2012] Standards for privacy ofindividually identifiable health information. Final Rule, 45 CFR parts 160 and 164. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf
3. Benitez, K.; Loukides, G.; Malin, BA. Beyond safe harbor: automatic discovery of health information de-identification policy alternatives. Paper presented at: The 1st ACM International Health Informatics Symposium; 2010.
4. Baumer D, Earp JB, Payton FC. Privacy of medical records: IT implications of HIPAA. ACM Computers and Society (SIGCAS). 2000; 30(4):40–47.
5. McGraw, D. [Accessed Feburary 20, 2012] Why the HIPAA privacy rules would not adequately protect personal health records: Center for Democracy and Technology (CDT) brief. 2008. http://www.cdt.org/brief/why-hipaa-privacy-rules-would-not-adequately-protect-personal-health-records
6. Foundation CH. National Consumer Health Privacy Survey 2005. 2005. p. 1-5.
7. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. Journal of the American Medical Informatics Association. 2012 (Epub ahead of print).
8. Tacconelli E. Systematic reviews: CRD’s guidance for undertaking reviews in health care. The Lancet Infectious Diseases. 2010; 10(4):226.
9. Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. Journal of the American Medical Informatics Association. 2011; 18(Suppl 1):103–108.
10. Vinterbo SA, Sarwate AD, Boxwala AA. Protecting count queries in study design. Journal of the American Medical Informatics Association. 2012 (Epub ahead of publish).
11. Fefferman NH, O’Neil EA, Naumova EN. Confidentiality and confidence: is data aggregation a means to achieve both? Journal of public health policy. 2005; 26:430–449. [PubMed: 16392743]
12. Matthews GJ, Harel O. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statistics Surveys. 2011; 5:1–29.
13. Fayyoumi E, Oommen BJ. A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. Software: Practice and Experience. 2010; 40:1161–1188.

14. Ohno-Machado L, Silveira P, Vinterbo S. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. International Journal of Medical Informatics. 2004; 73(7–8):599–606. [PubMed: 15246040]

15. Ohrn A, Ohno-Machado L. Using Boolean reasoning to anonymize databases. Artificial Intelligence in Medicine. 1999; 15(3):235–254. [PubMed: 10206109]

16. Jiang, X.; Cheng, S.; Ohno-Machado, L. Quantifying record-wise data privacy and data representativeness. Proceedings of the 2011 workshop on data mining for medicine and healthcare; San Diego, CA: ACM; 2011. p. 64-67.

17. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing. ACM Computing Surveys. 2010; 42:1–53.

18. Dwork C. Differential privacy: A survey of results. Theory and Applications of Models of Computation. 2008; 4978:1–19.

19. Bertino E, Lin D, Jiang W. A Survey of quantification of privacy preserving data mining algorithms. Privacy-Preserving Data Mining. 2008:183–205.

20. Zhao, Y.; Du, M.; Le, J.; Luo, Y. A survey on privacy preserving approaches in data publishing. Paper presented at: 2009 First International Workshop on Database Technology and Applications; 2009.

21. Sweeney L. *k*-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems. 2002; 10(5):557–570.

22. Sweeney, L. Datafly: a system for providing anonymity in medical data. Paper presented at: Eleventh International Conference on Database Securty XI: Status and Prospects; 1998; Chalkidiki, Greece.

23. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. *l*-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007; 1(1):3-es.

24. Li, N.; Li, T.; Venkatasubramanian, S. *t*-closeness: privacy beyond *k*-anonymity and *l*-diversity. Paper presented at: The 23rd International Conference on Data Engineering (ICDE); 2007.

25. Xiao, X.; Tao, Y. M-invariance: towards privacy preserving re-publication of dynamic datasets. Paper presented at: Proceedings of the 2007 ACM SIGMOD international conference on Management of data; 2007.

26. Cormode G, Srivastava D, Li N, Li T. Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data. Proc VLDB Endow. 2010; 3:1045–1056.

27. Ganta, SR.; Kasiviswanathan, SP.; Smith, A. Composition attacks and auxiliary information in data privacy. Paper presented at: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining -KDD ‘082008; New York, New York, USA.

28. Domingo-Ferrer, J.; Torra, V. A Critique of k-Anonymity and Some of Its Enhancements. IEEE; 2008.

29. Wong RC-W, Fu AW-C, Wang K, Pei J. Anonymization-based attacks in privacy-preserving data publishing. ACM Transactions On Database Systems. 2009; 34:8.

30. Wong RC-W, Fu AW-C, Wang K, Yu PS, Pei J. Can the Utility of Anonymized Data be Used for Privacy Breaches? ACM Transactions on Knowledge Discovery from Data. 2011; 5:1–24.

31. Sacharidis D, Mouratidis K, Papadias D. k-Anonymity in the Presence of External Databases. IEEE Transactions on Knowledge and Data Engineering. 2010; 22:392–403.

32. Cormode, G. Personal privacy vs population privacy: learning to attack anonymization. Paper presented at: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining; 2011; New York, NY, USA.

33. Nergiz, ME.; Atzori, M.; Clifton, C. Hiding the presence of individuals from shared databases. International Conference on Management of Data (SIGMOD); New York, NY, USA: ACM Press; 2007. p. 665-676.

34. Nergiz ME, Clifton C. δ-Presence without Complete World Knowledge. IEEE Transactions on Knowledge and Data Engineering. 2010; 22:868–883.

35. Muralidhar, K.; Sarathy, R. Does differential privacy protect terry gross’ privacy?. Paper presented at: Proceedings of the 2010 international conference on Privacy in statistical databases; 2010; Corfu, Greece.

36. Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering. 2002; 14:189–201.

37. Domingo-Ferrer J, Torra V. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Mining and Knowledge Discovery. 2005; 11:195–212.

38. Bayardo, RJ.; Agrawal, R. Data privacy through optimal *k*-anonymization. Paper presented at: The 21st International Conference on Data Engineering (ICDE); 2005.

39. Aggarwal G, Feder T, Kenthapadi K, et al. Anonymizing tables. Database Theory-ICDT 2005. 3363/20052005:246–258.

40. Chi-Wing, R.; Li, J.; Fu, AW-C.; Wang, K. (α, k)-anonymity. Paper presented at: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '062006; New York, NY, USA.

41. Ghinita G, Karras P, Kalnis P, Mamoulis N. A framework for efficient data anonymization under privacy and accuracy constraints. ACM Transactions on Database Systems. 2009; 34:1–47.

42. LeFevre, K.; DeWitt, DJ.; Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. Paper presented at: 22nd International Conference on Data Engineering (ICDE'06); 2006.

43. Friedman A, Wolff R, Schuster A. Providing k-anonymity in data mining. The VLDB Journal. 2007; 17:789–804.

44. LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. ACM Transactions on Database Systems. 2008; 33:1–47.

45. El Emam K, Dankar FK, Issa R, et al. A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association. 2009; 16:670–682. [PubMed: 19567795]

46. Mohammed N, Fung BCM, Hung PCK, Lee C-K. Centralized and Distributed Anonymization for High-Dimensional Healthcare Data. ACM Trans Knowl Discov Data. 2010; 4(4):18:11–18:33.

47. Kifer, D.; Gehrke, J. Injecting utility into anonymized datasets. Paper presented at: Proceedings of the 2006 ACM SIGMOD international conference on Management of data - SIGMOD '062006; New York, New York, USA.

48. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modeling Performance. Journal of the American Medical Informatics Association. 2002; 9(90061):115S–119.

49. Gouweleeuw J, Kooiman P, Willenborg L, De Wolf P. Post randomisation for statistical disclosure control: Theory and implementation. Journal of Offical Statistics. 1998; 14:463–478.

50. Vinterbo, SA.; Sarwate, AD.; Boxwala, A. Protecting count queries in cohort identification. Paper presented at: AMIA Summit on Clinical Research Informatics (CRI'11); 2011; San Francisco.

51. Barak, B.; Chaudhuri, K.; Dwork, C.; Kale, S.; Mcsherry, F.; Talwar, K. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. Paper presented at: Principles of Database Systems (PODS); 2007; Beijing.

52. Xiao X, Wang G, Gehrke J. Differential Privacy via Wavelet Transforms. Knowledge and Data Engineering, IEEE Transactions on. 2011; 23(8):1200–1214.

53. Mishra, DK. Tutorial: Secure Multiparty Computation for Cloud Computing Paradigm. 2010 Second International Conference on Computational Intelligence, Modelling and Simulation; Bali, Indonesia. 2010. p. xxiv-xxv.

54. Lindell Y, Pinkas B. Secure Multiparty Computation for Privacy-Preserving Data Mining. Journal of Privacy and Confidentiality. 2009; 1(1):59–98.

55. Clifton, C.; Kantarcioglu, M.; Vaidya, J. Privacy-preserving data mining. Vol. 180. New York, NY: Springer-Verlag; 2006.

56. Vaidya J, Yu H, Jiang X. Privacy-preserving SVM classification. Knowledge and Information Systems. 2008; 14(2):161–178.

57. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid LOgistic REgression (GLORE): Building Shared Models Without Sharing Data. Journal of American Medical Informatics Association. 2012 (Epub ahead of print).

58. Inan A, Kaya SV, Saygin Y, Savas E, Hintoglu AA, Levi A. Privacy preserving clustering on horizontally partitioned data. Data & Knowledge Engineering. 2007; 63(3):646–666.

59. Vaidya, J.; Clifton, CW. Privacy preserving association rule mining in vertically partitioned data. Paper presented at: Proceedings of the Eighth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining; 2002; Edmonton, Canada.

60. El Emam K, Hu J, Mercer J, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. Journal of the American Medical Informatics Association. 2011; 18(3):212–212. [PubMed: 21486880]

61. Xu, Z.; Yi, X. Classification of privacy-preserving distributed data mining protocols. Paper presented at: The Sixth International Conference on Digital Information Management; 2011.

62. Zhan J. Privacy-preserving collaborative data mining. IEEE Computational Intelligence Magazine. 2008; 3(2):31–41.

63. Zhang, Q.; Koudas, N.; Srivastava, D.; Yu, T. Aggregate query answering on anonymized tables. Paper presented at: The 23rd International Conference on Data Engineering (ICDE); 2007.

64. Nergiz ME, Tamersoy A, Saygin Y. Instant anonymization. ACM Transactions on Database Systems. 2011; 36(1):1–33.

65. Dinur, I.; Nissim, K. Revealing information while preserving privacy. Paper presented at: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; 2003; New York, NY, USA.

66. Blum, A.; Dwork, C.; McSherry, F.; Nissim, K. Practical privacy: the SuLQ framework. Paper presented at: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; 2005; New York, NY, USA.

67. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. Paper presented at: Theory of Cryptography Conference (TCC); 2006; New York, NY, USA.

68. Dwork C. Differential privacy. Automata, languages and programming. 2006; 4052/2006:1–12.

69. Roth, A.; Roughgarden, T. Proceedings of the 42nd ACM symposium on Theory of computing - STOC ‘102010;

70. McSherry F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. Communications of the ACM. 2010; 53(9):89–89.

71. Friedman, A.; Schuster, A. Data mining with differential privacy. Paper presented at: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ‘10); 2010; New York, NY, USA.

72. Roy, I.; Setty, STV.; Kilzer, A.; Shmatikov, V.; Witchel, E. Airavat: Security and privacy for MapReduce. Paper presented at: Proceedings of the 7th USENIX conference on Networked systems design and implementation; 2010.

73. Zhang N, Zhao W. Privacy-Preserving OLAP: An Information-Theoretic Approach. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(1):122–138.

74. Narayanan, A.; Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. Paper presented at: 2008 IEEE Symposium on Security and Privacy (sp 2008); 2008.

75. Hansell S. Removes Search Data On Vast Group Of Web Users. New York Times. 2006; 8:C4.

76. Lasko TA, Vinterbo SA. Spectral Anonymization of Data. IEEE transactions on knowledge and data engineering. 2010; 22:437–446. [PubMed: 21373375]

77. El Emam K, Dankar FK. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association. 2008; 15(5):627–627. [PubMed: 18579830]

78. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR). 2009; 42(4):1–53.

79. Fung BCM, Wang K, Wang L, Hung PCK. Privacy-preserving data publishing for cluster analysis. Data & Knowledge Engineering. 2009; 68(6):552–575.

80. Chaytor, R.; Wang, K.; Brantingham, P. Fine-Grain Perturbation for Privacy Preserving Data Publishing. Paper presented at: 2009 Ninth IEEE International Conference on Data Mining; 2009.

81. Wong R, Li J, Fu A, Wang K. (alpha, k)-anonymous data publishing. Journal of Intelligent Information Systems. 2009; 33(2):209–234.

82. Jin, X.; Zhang, M.; Zhang, N.; Das, G. Versatile publishing for privacy preservation. Paper presented at: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD); 2010; New York, NY, USA.

83. Arasu, A.; Kaushik, R.; Li, J. Data generation using declarative constraints. Paper presented at: Proceedings of the 2011 international conference on Management of data (SIGMOD 2011); 2011; New York, NY, USA.

84. Cano, I.; Torra, V. Generation of synthetic data by means of fuzzy c-Regression. Paper presented at: 2009 IEEE International Conference on Fuzzy Systems; 2009.

85. Domingo-Ferrer J, González-Nicolás Ú. Hybrid microdata using microaggregation. Information Sciences. 2010; 180:2834–2844.

86. Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. Very Large Database (VLDB) Endowment. 2010; 3(1–2):1021–1032.

87. Barak, B.; Chaudhuri, K.; Dwork, C.; Kale, S.; McSherry, F.; Talwar, K. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. Paper presented at: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; 2007; New York, NY, USA.

88. Kasiviswanathan, SP.; Rudelson, M.; Smith, A.; Ullman, J. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. Paper presented at: Proceedings of the 42nd ACM symposium on Theory of computing (STOC '10); 2010; New York, New York, USA.

89. Li, Y.; Zhang, Z.; Winslett, M.; Yang, Y. Compressive Mechanism: Utilizing sparse representation in differential privacy. Paper presented at: The Workshop on Privacy in the Electronic Society (WPES); 2011; Chicago, IL, USA.

90. Blum, A.; Ligett, K.; Roth, A. A learning theory approach to non-interactive database privacy. Paper presented at: Proceedings of the 40th annual ACM symposium on Theory of computing; 2008; New York, NY, USA.

91. Wasserman L, Zhou S. A statistical framework for differential privacy. Journal of the American Statistical Association. 2010; 105(489):375–389.

92. Mohammed, N.; Chen, R.; Fung, BCM.; Yu, PS. Differentially private data release for data mining. Paper presented at: International Conference on Knowledge Discovery and Data Mining (SIGKDD); 2011; San Diego, CA.

93. Machanavajjhala, A.; Kifer, D.; Abowd, J.; Gehrke, J.; Vilhuber, L. Privacy: Theory meets Practice on the Map. 2008 IEEE 24th International Conference on Data Engineering (ICDE); IEEE. 2008. p. 277-286.

94. Tao Y, Chen H, Xiao X, Zhou S, Zhang D. ANGEL: enhancing the utility of generalization for privacy preserving publication. IEEE Transactions on Knowledge and Data Engineering. 2009; 21(7):1073–1087.

95. Rastogi, V.; Suciu, D.; Hong, S. The boundary between privacy and utility in data publishing. Paper presented at: The 33rd International Conference on Very Large Data Bases (VLDB); 2007.

96. Goldberger, J.; Tassa, T. Efficient anonymizations with enhanced utility. Paper presented at: 2009 IEEE International Conference on Data Mining Workshops; 2009.

97. Domingo-Ferrer, J.; Rebollo-Monedero, D. Measuring risk and utility of anonymized data using information theory. Paper presented at: Proceedings of the 2009 EDBT/ICDT Workshops on - EDBT/ICDT '092009; New York, New York, USA.

98. Rebollo-Monedero D, Forné J, Soriano M. An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. Data & Knowledge Engineering. 2011; 70(10):892–921.

99. Malin, BA.; Sweeney, LA. Determining the identifiability of DNA database entries. AMIA Annual Symposium Proceedings; 2000. p. 537-541.

100. Malin BA. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. Journal of the American Medical Informatics Association. Jan-Feb;2005 12(1):28–34. [PubMed: 15492030]

101. Malin BA, Sweeney LA. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics. Jun; 2004 37(3):179–192. [PubMed: 15196482]

102. Malin BA, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. Hum Genet. Sep; 2011 130(3):383–392. [PubMed: 21739176]

103. Malin BA, Sweeney LA. Inferring genotype from clinical phenotype through a knowledge based algorithm. Pac Symp Biocomput. 2002:41–52. [PubMed: 11928494]

104. Chen, R.; Mohammed, N.; Fung, BCM.; Desai, BC.; Xiong, L. Publishing Set-Valued Data via Differential Privacy. The 37th International Conference on Very Large Data Bases; 2011. (in press)

105. Wang, K.; Xu, Y.; Wong, RC-W.; Fu, AW-C. Anonymizing temporal data. Paper presented at: International Conference on Data Mining; 2010.

106. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc. 2007; 14:574–580. [PubMed: 17823086]

107. Gardner, J.; Xiong, L.; Li, K.; Lu, JJ. HIDE: heterogeneous information DE-identification. Paper presented at: Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT '092009; New York, New York, USA.

108. Li, L.; Wang, JZ. DDIT - A tool for DICOM brain images de-Identification. The 5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE); Wuhan, China: IEEE; 2011. p. 1-4.

109. National Opinion Research Center. [Accessed 5/23, 2012] http://www.norc.org/Pages/default.aspx

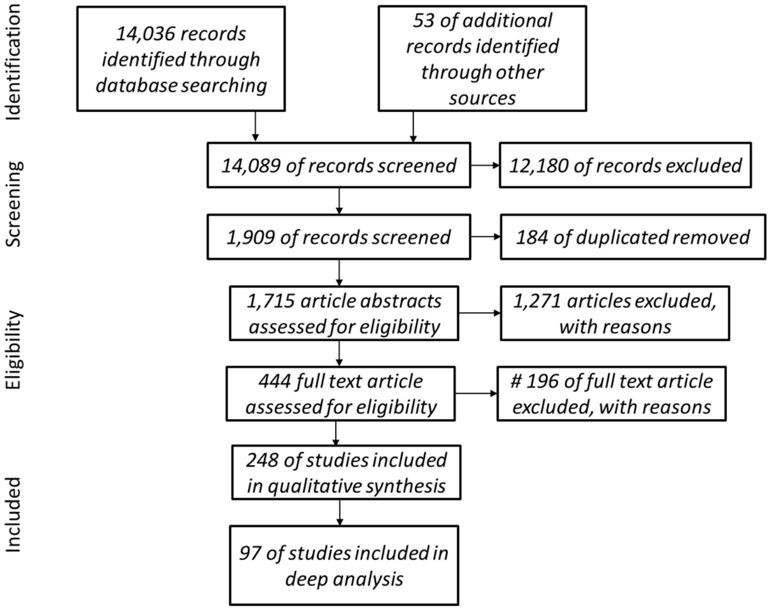
Identification
14,036 records identified through database searching
53 of additional records identified through other sources
Screening
14,089 of records screened
12,180 of records excluded
1,909 of records screened
184 of duplicated removed
Eligibility
1,715 article abstracts assessed for eligibility
1,271 articles excluded, with reasons
444 full text article assessed for eligibility
# 196 of full text article excluded, with reasons
Included
248 of studies included in qualitative synthesis
97 of studies included in deep analysis

Figure 1.

Workflow of the review process. 14,036 articles were retrieved from 7 major repositories and other resources. Screening based on title and keywords helped to remove 12,180 articles. Duplication check identified another 184 articles to be removed. The remaining 1,715 articles were reviewed and 1,271 were excluded based on the content in the abstract. The remaining 444 text were assessed by both reviewers, with 196 full articles excluded based on mutual consensus and arbitrage of a third party expert. The final list consisted of 248 articles with qualitative synthesis, from which 97 articles were selected for deep analysis.

Table 1

Description of the databases we queried, including major computer science, statistics, social science network, as well as medical literature repositories.

Web of Knowledge	Citation and indexing service provided by Thomson Reuters. Covers the sciences, social sciences, arts and humanities.
Social Sciences Research Network (SSRN)	Preprint and working paper repository hosted by a consortium of institutions. Covers law, economics, political science, policy, sociology, and related fields.
IEEExplore	Database of papers published by the Institute of Electrical and Electronics Engineers (IEEE). Covers computer science, engineering, and information management.
PubMed	Database and indexing service maintained by the United States National Library of Medicine (NLM) of the National Institutes of Health (NIH). Covers life sciences and biomedical topics.
JSTOR	Online journal storage system, JSTOR (short for Journal Storage), was founded in 1995. Covers over a thousand academic journals including mathematics, statistics, social sciences, and the humanities.
ACM Digital Library (ACM)	Database of papers published by Association for Computing Machinary (ACM) journal, newsletter articles and conference proceedings. Covers on computer and information science.
Arxiv	Preprint archive maintained by Cornell University Library. Covers mathematics, physics, astronomy, computer science, quantitative biology, statistics, and quantitative finance.

Table 2

Glossary and explanation for computer science terminology involved.

Glossary	Explanations
Suppression	Removing or eliminating certain features about the data prior to dissemination. For example, eliminate social security number.
Generalization	Transforming data into lower resolution (i.e., less detail). For example, generalize date of birth to year of birth, 5-digit to 3-digit ZIP code (e.g., 92130 to 921XX).
Perturbation	Producing specific outcomes with addition of noise. For example, adding random noise (e.g., +/−2) to the attribute age (e.g., age 40 gets transformed to age 42).
k-anonymity	A privacy criterion that specifies that each disclosed record has the exact same values for “k” people.
Differential privacy	A privacy criterion that quantifies the “indistinguishability” between databases that differ by at most one entry. This imposes an upper bound on the risk of inferences that an adversary can draw about the data, regardless of their background knowledge.
Contingency table	A matrix that represents the multivariate frequency distribution of variables.
Wavelet transform	A type of time-frequency transformation that represents a signal in terms of different scales.
Secure multiparty computation (SMC)	A subfield of cryptography that has the goal of enabling parities to jointly compute a function over inputs while preserving their privacy during information exchange.
Classification	The problem of identifying to which category (among several) a particular observation (record) belongs. For example, classifying patients into high/low-risk groups.
Clustering	Grouping a set of subjects so that subjects in the same group (i.e., cluster) are statistically more similar than those outside the group.
Association rule mining	Data mining methodology to reveal interesting relations between variables in large databases.
MapReduce	A programming model to handle large data sets though decomposing tasks into parallel distributed programs.
Count query	A query that returns the number of rows satisfying selection criteria.