# **Appendix A**

## **Statutory definitions for de-identifying phi**

**HIPAA defines two approaches for de-identification in CFR 45 §164.514. The first is expert determination, in which an expert (defined as an "individual with appropriate knowledge of and experience") applies "generally accepted statistical and scientific principles and methods" to ensure that there is "very little risk" that the "anticipated recipient" of the data would be able to re-identify an individual. The rule does not specify how risk should be quantified, but in practice the expectation is that the expert and data custodians will collaboratively develop a risk mitigation plan. Four major principles (reliability, data source availability, distinguishability, and access risk) serve as the basis for the determination of health information identifiability. Experts may also design mechanisms to allow the re-identification of de-identified information by authorized parties. Such an approach could be implemented using cryptographic methods, but cannot use auxiliary information about the individuals in the data set.**

**The second approach defined by HIPAA is known as the Safe Harbor method. To meet the statutory definitions, 18 types of identifiers regarding individual or relatives, employers, or household members of the individual should be carefully handled. The following list of identifiers are extracted from latest guidance on methods for de-identifying protected health information (PHI)[[1]](#footnote-1)\* from the Department of Health and Human Services (HHS).**

1. **Name**
2. **Geographic information that is smaller than a state (i.e., street address, city, county, precinct, ZIP code, or equivalent geocodes). If the area formed by merging all ZIP codes with the same first three digits has more than 20,000 people, then the first three digits of the ZIP code may be released, otherwise the first three digits should be set to 000.**
3. **All elements of dates (except years) should be excluded and all ages above 89 should be aggregated into a single category of ≥90**
4. **Telephone numbers**
5. **Fax numbers**
6. **Email addresses**
7. **Social security numbers**
8. **Medical record numbers**
9. **Health plan beneficiary numbers**
10. **Account numbers**
11. **Certificate / license numbers**
12. **Vehicle identifier and series numbers, including license plate numbers**
13. **Device identifiers and serial numbers**
14. **Web universal resources locators (URLs)**
15. **Internet protocol (IP) addresses**
16. **Biometric identifiers, including finger and voice prints**
17. **Full-face photographs and any comparable images**
18. **Any other unique identifying number, characteristic, or code**

**Situations involving unstructured documents (e.g., free text in doctor's notes) pose challenges because it is difficult to automatically find the unique identifying characteristics mentioned in item 18.**

**After the data have been de-identified using the safe harbor method, the privacy rules put no limitation on how they can be shared or disseminated. However, data custodians may still ask the recipient of de-identified data to sign a data use agreement (i.e., prohibiting the re-identification to prevent inappropriate information leakage) based on their own judgment.**

## Prior Work and Contribution

As noted by Golub and Fontanarosa[1](#_ENREF_1), “precise descriptions and definitions of CER and the inferences that can be drawn” are still subject to interpretation. Existing CER studies include both large multi-site randomized clinical trials as well as purely observational studies, and span many areas of medicine. Examples include topics as diverse as the effectiveness of antidepressants[2](#_ENREF_2), improving patient literacy[3](#_ENREF_3), radiology interventions[2](#_ENREF_2), and helicopter versus ground transport for trauma patients[4](#_ENREF_4). The focus of our study for data sharing is to categorize future CER studies by the structure of their data sharing arrangements and to identify current data sharing approaches that can support such CER studies.

Privacy has long been an object of study within the official statistics community[5-7](#_ENREF_5)[11](#_ENREF_11), [12](#_ENREF_12), with a focus on multiple imputation[8](#_ENREF_8) and synthetic data[9](#_ENREF_9). A recent survey that covers official statistics as well as some of the computer science literature can be found in Matthews and Harel[10](#_ENREF_10). Privacy technologies have also been surveyed in the computer science literature[11-17](#_ENREF_11). A recent survey of different privacy definitions, attack models, and privacy-preserving data manipulations can be found in a review by Fung et al.[18](#_ENREF_18), which is more comprehensive than earlier surveys[19-21](#_ENREF_19). In the medical context, much of the existing work has focused on disseminating data for public health purposes[22](#_ENREF_22),[23](#_ENREF_23). Leino-Kilpi surveyed privacy concepts in the context of nursing[24](#_ENREF_24). El-Emam and Fineberg have a high-level overview of approaches to de-identifying personal health information (PHI) for secondary purposes[25](#_ENREF_25). Our work builds on these surveys butdiffers from them in two respects. First, we focus onthe potential of existing technologies to apply to CER. We adopt a systematic review methodology to identify and evaluate tradeoffs between different technologies. We then identified several data sharing arrangements that may arise in future CER studies and classify the literature accordingly.

# Details of the review process

We split the task of querying the databases. Because the queries were purposely quite broad, our initial searches returned a very large number of responses. We manually filtered these results based on the title (**Pass I**) to determine if they were relevant to data sharing for CER. Many papers were not relevant, because words such as "re-identification" and "control" have different meanings in different research communities, and because many articles focused on policy or other issues related to privacy but not privacy technologies. We then read the abstracts of the remaining papers (**Pass II**) to classify papers as relevant or irrelevant. Papers were deemed relevant if they contained algorithms useful for CER and used a quantifiable definition of privacy. Papers were excluded if they were primarily encryption-based methods for communication or focused on security aspects of data sharing rather than privacy and confidentiality of the data. The focus on algorithms excluded many papers on cryptographic approaches to privacy, studies of privacy for location-based services in mobile devices, and other topics outside the scope of CER and healthcare. With these filtering guidelines, there were 444 papers left for reviewing in **Pass III**. Two authors independently reviewed the remaining list and assigned them to one of three categories. The “review” category contained papers which should undergo a full text review. The “relevant” category contained papers that were relevant for historical or contextual purposes, but did not directly provide privacy-preserving methods for the CER applications discussed in this paper. The “no-cite” category contained papers which were subsumed by earlier work or which, on further review, did not pertain to CER. Conflicts in excluding papers were resolved by either discussion among XJ and AS or referred to Dr. Robert El-Kareh, UCSD, for arbitrage. Of the 444 papers, 248 were tagged as “relevant” and 97 were tagged as “cite”. The full texts of these 97 papers were reviewed in more detail to generate results and recommendations in this review. An overview of the reviewing process is illustrated in Figure A1.The detailed inclusion, exclusion, and tagging criteria are elaborated later in this section.



**Figure A1:** Overview of the review process. Three steps of filtering are conducted based on title, abstract, and full text. The first two steps were conducted jointly by reviewers to quickly remove irrelevant studies. The last step was conducted independently by each reviewer and conflicts were resolved through discussion or arbitrage by a third party expert. The remaining articles entered the final analysis stage for this study.

# Inclusion and exclusion criteria (pass I and II)

**Table A2:** Inclusion and exclusion criteria in the first pass (based on title and keywords).

|  |  |
| --- | --- |
| **INCLUDE if:** | **EXCLUDE if:** |
| * contains quantifiable privacy metric in the abstract | * deals only on policy |
| * mentions a specific privacy-preserving statistical algorithm | * only mentions privacy as a factor to be considered |
| * mentions CER and privacy together | * is about location privacy or traffic data |
| * introduces techniques for privacy and | * is about social networks |
| security problems | * is about image privacy/face de-identification |
|  | * only deals with encryption methods (security, not privacy) |
|  | * only mentions data use agreements |
|  | * is about architectures for security systems |

**Table A3:** Inclusion and exclusion criteria in the second pass (based on title and key words).

|  |  |
| --- | --- |
| **INCLUDE if:** | **EXCLUDE if:** |
| * is relevant to CER and privacy | * is about secure multi-party computing |
| * has a quantifiable definition of privacy | * is about encryption frameworks |
|  | * is about secure systems/architectures |

# Tag the union of separately filtered papers



**Figure A2:** Tags assigned to the union of filtered papers in pass II. These tags include 6 major categories of characteristics, each comprised of a list of subcategories to capture various fine-grained aspects.

We excluded a total of 196 articles in our last stage of filtering based on full text. Table A2 summarizes the reasons of such exclusion,

**Table A2:** The reason of exclusion at the second pass of eligibility filtering.

|  |  |
| --- | --- |
| **Reason to exclude** | **Number of papers** |
| Approximation / heuristics | 18 |
| Specific data mining tasks not relevant to CER | 43 |
| Small derivatives of existing techniques | 42 |
| Empirical / regional studies | 5 |
| Algorithms complexity | 17 |
| Secure multiparty computation theory | 16 |
| Cryptography theory | 38 |
| Irrelevant data format | 8 |
| Various ad-hoc attack models | 15 |
| Overlaps | -6 |
| Subtotal | 196 |

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. Table A3 shows the queries we made on these databases and the number of returned entries. Because the queries were purposely broad, they sometimes returned a very large number of documents. In the first round, the responses were manually filtered based on the title to determine if they were relevant to data sharing for CER. A large number of papers were irrelevant; reasons for this included alternative meanings of terms such as “re-identification” or “control” in the research literature, a focus on healthcare policy rather than technologies, purely theoretical investigations, or applications to non-CER problems such as location privacy or social networking. The reviewers then read the abstracts of the remaining 1,715 articles (after removing duplicates), yielding 444 papers whose text was reviewed in detail. Among these, 97 articles entered the final stage for synthesis because they were directly relevant and contained quantifiable measures for privacy.

**Table A3:** Queries for documents on patient privacy and electronic informed consent systems. Relevant documents were determined based on abstract review.

|  |  |  |  |
| --- | --- | --- | --- |
| **Patient Privacy and Electronic Informed Consent Systems** | | | |
| **Database** | **Search Terms** | **# Retrieved** | **# Relevant** |
| **Medline** | "Confidentiality" AND "privacy" OR "De-identification" OR "Disclosure control" AND "privacy" OR "Identifiability" OR "Re-identification" | 5.366 | 198 |
| **Web of Knowledge** | "re-identification" OR "privacy-preserving" OR "disclosure control" AND "privacy" OR "de-identification" OR "anonymization" | 1,454 | 375 |
| **Social Sciences Research Network** | “privacy-preserving" OR "anonymity" OR "disclosure control" OR "privacy-preserving" OR "de-identification" AND "privacy" OR "anonymization" OR "privacy" AND "confidentiality" OR "re-identification" AND "privacy" | 746 | 13 |
| **IEEExplore** | "privacy" AND "confidentiality" OR "privacy-preserving" OR  "anonymity"+"identification" OR "anonymization" OR "de-identification" OR "re-identification" OR "disclosure control" OR "data privacy" AND "anonymity” | 2,900 | 575 |
| **JSTOR** | "Confidentiality+privacy" OR "De-identification" OR "Disclosure control" AND "privacy" OR "Identifiability" OR "Re-identification" | 74 | 16 |
| **ACM Digital Library** | "Confidentiality" AND "privacy" OR "De-identification" OR "Disclosure control" AND "privacy" OR "Identifiability" AND "privacy" OR "Re-identification" | 1,467 | 611 |
| **Arxiv** | "Confidentiality" AND "privacy" OR "De-identification" OR "Disclosure control" AND "privacy" OR "Identifiability" AND "privacy" OR "Re-identification" | 2,900 | 575 |
| Subtotal | | 14,036 | 1,858 |
| Duplicates | |  | -184 |
| Contributed citation from other sources | |  | +51 |
| **Total** | |  | **1,715** |

References:

**1.** Golub RM, Fontanarosa PB. Comparative Effectiveness Research : Relative Successes. *JAMA: The Journal of the American Medical Association.* 2012;307(15):1643-1645.

**2.** Gartlehner G, Hansen RA, Thieda P, et al. *Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression*: AHRQ;2007.

**3.** Dewalt DA, Schillinger D, Ruo B, et al. A Multisite Randomized Trial of a Single- versus Multi-Session Literacy Sensitive Self-Care Intervention for Patients with Heart Failure. *Circulation.* 2012.

**4.** Galvagno Jr SM, Haut ER, Zafar SN, et al. Association Between Helicopter vs Ground Emergency Medical Services and Survival for Adults With Major Trauma. *Journal of the American Medical Association (JAMA).* 2012;307(15):1602-1610.

**5.** Adam NR, Worthmann JC. Security-control methods for statisti- cal databases: a comparative study. *ACM Computing Surveys.* 1989;21(4):515-556.

**6.** Domingo-ferrer J. A Quantitative Comparison of Disclosure Control Methods for Microdata. *Confidentiality disclosure and data access Theory and practical applications for statistical agencies.* 2001:113-135.

**7.** Moore R. Controlled Data Swapping Techniques for Masking Public Use Data Sets. *U.S. Bureau of the Census, Statistical Research Division Report rr96/04* 1995; <http://www.census.gov/srd/www/byyear.html>.

**8.** Raghunathan TE, Reiter JP, Rubin DB. Multiple Imputation for Statistical Disclosure Limitation. *Journal of Offical Statistics.* 2003;19(1):1-16.

**9.** Winkler WE. Masking and re-identification methods for public-use microdata: Overview and research problems. *Proceedings of Privacy In Statistical Databases.* 2004;3050:231-246.

**10.** 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.

**11.** Cormode G, Srivastava D. Anonymized Data: Generation, models, usage. Paper presented at: 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)2010.

**12.** Agrawal R, Srikant R. Privacy-preserving data mining. *ACM Sigmod Record.* 2000;29(2):439-450.

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

**14.** 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 Applications2009.

**15.** Aggarwal CC, Yu PS. *Privacy-preserving data mining: models and algorithms*. New York, NY: Springer; 2008.

**16.** Nayak G, Devi S. A Survey on Privacy Preserving Data Mining Approaches and Techniques. *International Journal of Engineering Science.* 2011;3(3):2127-2133.

**17.** Luo Y, Zhao Y, Le J. A Survey on the Privacy Preserving Algorithm of Association Rule Mining. *2009 Second International Symposium on Electronic Commerce and Security.* 2009;2:241-245.

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

**19.** Gehrke J. Models and Methods for Privacy-Preserving Data Analysis and Publishing. Paper presented at: 22nd International Conference on Data Engineering (ICDE'06)2006.

**20.** Gehrke J. Models and methods for privacy-preserving data publishing and analysis. Paper presented at: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '052005; New York, NY, USA.

**21.** Evfimievski A. Randomization in privacy preserving data mining. *ACM SIGKDD Explorations Newsletter.* 2002;4:43-48.

**22.** 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.

**23.** El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research.* 2006;8:e28.

**24.** Leino-Kilpi H, Välimäki M, Dassen T, et al. Privacy:a review of the literature. *International Journal of Nursing Studies.* 2001;38:663–671.

**25.** Emam KE, Fineberg A. An Overview of Techniques for De-Identifying Personal Health Information (Working Paper Series). 2009; <http://ssrn.com/abstract=1456490>.

1. \*<http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html> [↑](#footnote-ref-1)