Published in final edited form as:
Med Care. 2012 July ; 50(Suppl): S82–101. doi:10.1097/MLR.0b013e3182585355.

Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies

Clete A. Kushida, M.D., Ph.D.[1], **Deborah A. Nichols, M.S.**[1], **Rik Jadrnicek**[2], **Ric Miller**[3], **James K. Walsh, Ph.D.**[4], and **Kara Griffin**[4]

[1]Stanford Sleep Medicine Center, Redwood City, CA

[2]Microflow DBMS Inc., Aalapapa Drive, Kailua, HI

[3]Microflow DBMS Inc., Sausalito, CA

[4]Sleep Medicine and Research Center, Chesterfield, MO

Abstract

Background: De-identification and anonymization are strategies that are used to remove patient identifiers in electronic health record (EHR) data. The use of these strategies in multicenter research studies is paramount in importance, given the need to share EHR data across multiple environments and institutions while safeguarding patient privacy.

Methods: Systematic literature search using keywords of deidentify, de-identify, deidentification, de-identification, anonymize, anonymization, data scrubbing, and text scrubbing. Search was conducted up to June 30, 2011 and involved 6 different common literature databases. A total of 1,798 prospective citations were identified, and 94 full-text articles met the criteria for review and the corresponding articles were obtained. Search results were supplemented by review of 26 additional full-text articles; a total of 120 full-text articles were reviewed.

Results: A final sample of 45 articles met inclusion criteria for review and discussion. Articles were grouped into text, images, and biological sample categories. For text-based strategies, the approaches were segregated into heuristic, lexical, and pattern-based systems vs. statistical learning-based systems. For images, approaches that de-identified photographic facial images and magnetic resonance image data were described. For biological samples, approaches that managed the identifiers linked with these samples were discussed, particularly with respect to meeting the anonymization requirements needed for IRB exemption under the Common Rule.

Conclusions: Current de-identification strategies have their limitations, and statistical learning-based systems have distinct advantages over other approaches for the de-identification of free text. True anonymization is challenging, and further work is needed in the areas of de-identification of data sets and protection of genetic information.

Corresponding Author Contact: Clete A. Kushida, M.D., Ph.D., Stanford Sleep Medicine Center, 450 Broadway Street, MC 5704, Pavilion C, 2nd Floor, Redwood City, CA 94063-5704, T: 650-721-7560, F: 650-721-3465, clete@stanford.edu.

Keywords

de-identification; anonymization; electronic health record

INTRODUCTION

The use of medical records and human tissues in biomedical research in the U.S. is covered under the Standards for Privacy of Individually Identifiable Health Information (usually referred to as the Privacy Rule), and The Common Rule. In response to a congressional mandate in the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Department of Health and Human Services (HHS) issued the HIPAA Privacy Rule regulations in December 2000. The Privacy Rule permits covered entities (i.e., health plans, health care clearinghouses, or health care providers who transmit health information in electronic form in connection with a transaction for which HHS has adopted standards) to use and disclose data that have been removed of patient identifiers without obtaining an authorization and without further restrictions on use or disclosure because data removed of these identifiers are no longer protected health information (PHI) and, therefore, are not subject to the Privacy Rule. There are 18 “safe harbor” data identifiers under the Privacy Rule that constitute the minimal set of removed identifiers. The Common Rule sets the basic principles for protecting patients from research risks, using human tissues in support of medical research, and guiding the activities of Institutional Review Boards.

The use of data removed of patient identifiers is one of three current options available to investigators desiring to use medical data in research, besides obtaining informed consent from their patients or a waiver of informed consent from their institutional review board (IRB). The processes by which a data custodian prepares, manages, and distributes a data set that does not contain individually identifiable information to a data recipient is referred to as de-identification or anonymization (Table 1). *De-identification* of medical record data refers to the removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data. Although a de-identified data set may contain an encrypted patient identifier with which authorized individuals could re-link a patient with his or her data set, this data set must not contain data that will allow an unauthorized individual to infer a patient’s identity from the existing data elements. *Anonymization* refers to the irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link. With IRB approval, an individual can be re-identified from a de-identified record, but this is not the case for an anonymized record. However, it is of concern that de-identified and even anonymized patient data sets could potentially be re-identified. For example, based on experiments using 1990 U.S. Census summary data, Sweeney reports that 87% of the population in the United States reported characteristics that likely made them unique based only on 5-digit zip code, gender, and date of birth.(1)

As the use of electronic health records (EHRs) has progressively increased, concerns have been raised about their utility to fundamentally improve the quality of patient care and the threat of unauthorized disclosure of PHI either unintentionally or by identity theft.

Additionally, biomedical research is becoming increasingly dependent on the access, sharing, and management of EHR among clinical and research centers, especially those involved in observational and multicenter research studies, in particular comparative effectiveness research (CER). The American Recovery and Reinvestment Act (ARRA) of 2009 included $1.1 billion for CER to support research assessing the comparative effectiveness of health care treatments and strategies. As observed by others, with this influx of support for CER combined with the digitization of medical and biological information, we appear to be closer than ever to providing healthcare's stakeholders with much needed evidence of the comparative effectiveness of treatments. The use of de-identification and anonymization strategies in multicenter research studies, and especially CER, is critically important, since this type of research typically involves large-scale environments encompassing multiple institutions, and these strategies provide a scalable way for sharing medical information in these environments while also protecting patient privacy. Below we review published strategies and techniques specifically developed for de-identification and anonymization of EHRs and comment on their strengths and limitations.

METHODS

Database Sources for Literature Search (Figure 1)

- BIOSIS Previews (via Thomson Reuters Institute for Scientific Information [ISI] Web of Knowledge, 1926-present)
- CINAHL (Cumulative Index to Nursing and Allied Health Literature, via EBSCOhost, 1937-present)
- Inspec (via Thomson Reuters ISI Web of Knowledge, 1898-present)
- MEDLINE (Medical Literature Analysis and Retrieval System Online, 1950-present)
- SciVerse Scopus (1823-present)
- Web of Science (via Thomson Reuters ISI Web of Knowledge, 1898-present)

Key Words

The following keywords were used to identify prospective articles: deidentify, de-identify, deidentification, de-identification, anonymize, anonymization, data scrubbing, and text scrubbing. In three of the databases (BIOSIS Previews, Inspec, Web of Science), a wildcard (e.g., *,$) was placed in substitution of the hyphen in "de-identify" and "de-identification" to allow the database search tool to correctly recognize these keywords.

Search Criteria and Strategy

Articles were included if they were published up to June 30, 2011 and there was no restriction on earliest date of publication (i.e., earliest date obtained in search was 1996). Through the combined database search, 1,798 prospective citations were identified (Figure 1). Duplicate citations, citations that were not relevant to the topic, and citations for non-relevant article types (e.g., reviews, opinions, editorials, or commentaries) were excluded. Abstracts of citations that appeared relevant were reviewed, and those that were either not relevant to the topic or outside of the medical records domain were excluded. In addition, abstracts were excluded if the corresponding article was written in a language other than English, but there was no exclusion for research conducted in other countries and it should be recognized that their strategies do not necessarily need to conform to HIPAA and other U.S. rules.

A total of 94 full-text articles met the criteria for review and the corresponding articles were obtained; search results were supplemented by review of 26 additional full-text articles through extraction of relevant citations from the references of articles obtained through the searches. This produced a total of 120 full-text articles reviewed; 75 articles were excluded since they were not relevant to the topic of de-identification or anonymization strategies or because the given strategy lacked sufficient detail to understand or interpret it. Articles focused specifically on anonymization strategies were included only if they had been tested on a medical population; articles encompassing anonymization models (e.g., *k*-anonymity), primarily based on generalization methods, were not included in this document. The chair of the writing group (CK) conducted the above process; however, five other members of the writing group (JW, RJ, RM, DN, KG) independently reviewed the 120 full-text articles obtained after the abstracts review. Differences between the reviewers' judgments regarding inclusion or exclusion of articles were resolved by discussion; consensus was required from all six reviewers. Thus, the full text of 120 articles were reviewed and resulted in a final sample of 45 articles that met inclusion criteria for review and discussion in this document.

CASE EXAMPLES

Text

Manual de-identification of PHI from free text in EHR can be tedious, costly, time-consuming, inaccurate, and unreliable. Neamatullah reported that resident clinicians could de-identify at a rate of about 18,000 words or 90 incidents of PHI per hour.(2) Automated de-identification approaches to de-identify free text in EHR have been developed, and there are various ways of classifying these approaches. One method is to segregate heuristic, lexical, and pattern-based systems vs. statistical learning-based systems. The heuristic, lexical, and pattern-based systems rely on matching precompiled, manually-constructed sets of word lists, dictionaries, regular expressions, and heuristics to remove occurrences found in the free text portions. Statistical learning-based systems typically use an annotated training set of examples to search for a statistical pattern of specific features, learning how to identify PHI from the data itself.

An example of a heuristic, lexical, and pattern-based system is *deid*, which is an automated Perl-based de-identification software package that uses lexical look-up tables, regular expressions, and simple heuristics to locate both HIPAA PHI, and an extended PHI data set that includes doctors' names and years extracted from dates (Table 2).(2) The de-identification process involves scanning the free-text EHR (e.g., nursing notes, discharge summaries, X-ray reports) line-by-line, and dividing them into individual words; PHI is then identified using dictionary-based look-ups and pattern matching using regular expressions. Simple heuristics qualify or reject ambiguous terms as PHI. The final step in the process is the replacement of each instance of PHI with a tag to indicate its corresponding category.

This system had equal or better performance than manual de-identification, with an average recall performance of 96.7% on a gold-standard corpus of re-identified nursing notes. This system has also been adapted for regional use in Sweden (3) (4), France,(5) and Ontario, Canada(6). Additionally, other heuristic, lexical, and pattern-based systems use strategies such as algorithms employing rules, dictionaries, look-up tables, word list-matching, and pattern-matching either singly or in combination;(7–19) semantic category recognition approaches;(20) numerous detection algorithms competing in parallel to recognize a specific entity (e.g., fields such as first name, last name, street address, date);(21) natural language processing (NLP) tools to identify and remove PHI;(22–24) and disambiguating clinical text containing jargon and acronyms using rules that analyze surrounding words followed by data scrubbing.(25)

By contrast, an example of a statistical learning-based system is *Stat De-id*, which uses support vector machines (SVM) to identify the class of individual words as one of 8 categories (doctor, location, phone, address, patient, ID, hospital, or non-PHI), and uses features of the target, as well as features of surrounding words in order to capture contextual relationships.(26) Syntactic information extracted from the *Link Grammar Parser* (27) was used to create a representation of local context, which aids de-identification despite PHI including out-of-vocabulary words, and even when PHI are ambiguous with non-PHI within the same corpus. Semantic information from medical dictionaries was used to complement the syntactic information. *Stat De-id* was evaluated using 889 discharge summaries, and the system achieved an F-measure of 98% with a precision of 99% and a recall of 97% (average of all PHI types). Other statistical learning-based systems contain unique features such as an algorithm that estimates the probability that an assigned candidate patient name reference satisfies a set of semantic selection restrictions;(28) an SVM that was trained with syntactic and lexical contextual clues, as well as ontological information from Unified Medical Language System (UMLS);(29) a Named Entity Recognition (NER) system based on Conditional Random Fields (CRF) (i.e., discriminative probabilistic model that is shown to be effective in labeling natural language) for extracting identifying and sensitive attributes;(30) and a toolkit that combines a Web-based graphical annotation tool with modules for document processing, redaction and resynthesis, and engine, and evaluation of the performance of the PHI classifier.(31, 32).

The American Medical Informatics Association (AMIA) sponsored an automated de-identification challenge in 2007 as part of the i2b2 (Informatics for Integrating Biology and the Bedside) project.(33) A corpus of medical discharge summaries (training set of 669 reports and test set of 220 reports) was used by seven teams of investigators using heuristic and/or statistical de-identification strategies. Statistical learning systems using rule templates as features performed best, with the best performing de-identification system used two existing toolkits (*Carafe* and *LingPipe*) for NER complemented with regular expressions that could capture the more standardized PHI, followed by hybrid systems of rules then machine learning(34). The seven systems had recall rates of 80–96% with specificities of 83–97%.(33)

Images

HIPAA guidelines require that for data to qualify as sharable under the Privacy Rule “safe harbor” regulations, one of the identifiers that must be removed is “full facial photographic images and any comparable images.” However, there is a need for maxillofacial surgeons and other clinicians to view and share these images to improve treatment and for educational purposes. To address this challenge, a method that uses blended facial composites to de-identify photographic facial images was developed.(35) These composites were created from a library of frontal facial images of white men obtained from a secure image repository. At least 2 original facial images were selected, based on a subjective analysis of facial characteristics and an expected best fit, and these images were combined into a single facial composite in 10 to 15 minutes. Participants strongly agreed (83%) or agreed (17%) that the composites were clinically realistic patient images, and 83% rated the composites as more effective at de-identification than traditional methods. However, there is a concern regarding the degree to which the composite alters the original image thereby decreasing its realism and clinical usefulness.

Aside from full-face images, facial features are also important when recognizing a familiar individual, so researchers have the added complication of removing identifying facial features from morphometric scans due to concerns that these images can be used to identify individuals. The significance of this problem is exemplified by one study that used 3-D volumetric reconstructions of magnetic resonance (MR) image data, in which participants were requested to pair these reconstructions with one of 40 photographs; 40% of the participants were able to successfully match photographs with MR images.(36) The need to de-identify facial features, however, must be balanced against methods that affect either the quality of the image or the removal or distortion of brain tissue and/or other anatomical features under study. Automated skull-stripping algorithms (i.e., for removal of non-cerebral tissue in T1-weighted MR brain images) are commercially available that could be used for de-identification purposes, but subject population and scanner performance during data acquisition may degrade the image quality of the region of interest and prove unreliable for multicenter studies that require large-scale, automated de-identified processes for image sharing.(37) A newer approach uses an automated “defacing” algorithm that uses models of non-brain structures for removing only identifiable facial features from MR volumes. This defacing algorithm did an effective job of removing facial features without sacrificing brain tissue, could be performed relatively quickly (approximately 25 min on a dataset of 342), did not interfere with subsequent data processing, and in some cases, improved the quality of subsequent automated skull-stripping by removing more non-brain tissue.

Besides full-face de-identification, skull-stripping, and defacing algorithms, other investigators have developed approaches for Digital Imaging and Communications in Medicine (DICOM) data, which involve either substituting a new de-identified DICOM header;(38) altering PHI instances (e.g., removing PHI by setting the value to null or emptying their content, substituting dummy or new, de-identified values);(39, 40) using reversible anonymization by encryption and early removal of identifying data;(39) or using threshold detection in an algorithm to calculate region variances of intensity values for each pixel in the image to separate PHI and annotations from background and anatomic structures.(41) Unfortunately, quantitative data on the performance of these algorithms were not reported in these studies, so it is difficult to draw conclusions on their efficacy in de-identifying these data.

Biological Samples

The management of identifying data linked to biological samples is becoming increasingly important as multicenter clinical trials and CER require methods to process samples obtained from diverse locations and stored in biobanks. A method of de-identification of biological samples for genetic research uses a third-party encryption method developed by deCODE genetics and the Data Protection Commission of Iceland (DPC).(42) Physicians create a population-based list of the patients with a particular disease for the DPC; this list is reversibly encrypted by the DPC such that each social security number (SS) is converted to an alphabet-derived character string (PN). The laboratory receives the PN list and phenotype classification and compares the list against an encrypted population-based genealogy database, all prepared by the DPC. A list of patients of interest for a genetic study is sent to the DPC, and the list is decoded. Patients are contacted by physicians, and those willing to participate have their blood drawn and placed in tubes with a temporary-coded sample number (SN). The SN is scanned into a computer and the patient's SS is keyed in the presence of the patient, establishing a SS-SN link. The DPC officer encrypts the list of SS-SN to PN-SN, establishing the link of SN with PN. The PN-SN list is sent on a sealed computer diskette along with the blood to the laboratory. At the laboratory, the blood tubes are scanned into a sample storage program and the temporary SN is replaced by relabeling with an in-house sample number (iSN), and it will remain with the DNA sample isolated from the blood and is directly linked by the storage program to the PN used to label individuals within the database. Thus, the only connection between samples or data in the laboratory and the patient or data on the clinical side is through the PN and the DPC, the sole keeper of the encryption code. This third-party encryption method takes about 3 hours for a list of 1000 individuals to be encrypted or decrypted, and the cost of the system is related to the labor of the actual encryption and decryption by a trusted third party or representative.(42)

Furuta reported a different approach for de-identifying the blood samples in their bio-repository.(43) This repository contains approximately 250,000 samples with an average influx of 90,000 samples per year of which approximately 80% need to be de-identified. The de-identification process involves the transfer of samples from clinical identifier-tagged tubes collected from various clinical test settings to new repository identifier-tagged tubes without clinical patient identifiers; these samples are then transferred to the repository for final storage at −20°C. This process differs from data scrubbing patient identifiers on the physical sample since a tube transfer procedure is used for sample de-identification. However, this is a manual de-identification procedure that is subject to human error. To minimize such risk, the investigators manage a small number of tubes (e.g., 12 tubes) in each maneuver and have different personnel verify tube identifiers.

Both of these approaches for managing the identifiers linked with biological samples, as well as other approaches reported by Hara (44) and Roden (17), do not appear to be anonymization strategies, although the description provided by Furuta (43) does not specifically indicate whether the samples can be re-identified. Nevertheless, it appears that all of these strategies could be modified so that the samples are anonymized rather than simply de-identified. This is especially important for research conducted in the United States as tissue-related research activities may require IRB approval and data use agreements despite the HIPAA exemption for record de-identification, since these activities are subsumed under the Common Rule and de-identification may not meet the anonymization requirements needed for IRB exemption under this rule.

DISCUSSION

The review of these strategies uncovers key questions:

Are current de-identification strategies effective?

There is no question that current de-identification strategies have impressive recall and precision rates. However, no existing system is perfect, and there is the possibility that certain PHI will not be de-identified. However, all instances of PHI are not equal, with some identifiers (e.g., name) more critical than others. Additionally, limitations of many current systems include an inability to detect misspellings, typographical errors, and proper names that share characteristics with non-PHI (e.g., the family name "Black"); restrictions in managing only certain types of data (e.g., discharge summaries); algorithms that are not designed to handle diverse PHI (e.g., hard-coded or embedded PHI in device-generated output files); and difficulty in compensating for regional or geographic variation in nomenclature. The challenge is to balance the levels of de-identification that are acceptable to the patients, research participants, clinicians, researchers, institutions, and federal requirements (all of which may not necessarily be similar) with operational factors related to time, cost, and labor.

Which strategies are best?

In the case of de-identification of free text, both heuristic, lexical, and pattern-based systems vs. statistical learning-based systems have their advantages and disadvantages. For the former, studies evaluating these systems have reported good performance (especially precision) but experienced domain experts must spend significant amounts of time and effort developing, organizing, maintaining, and extending the rules, which likely need to be modified for different data sets. For statistical learning-based systems, they are able to be used "out of the box" with minimal redevelopment time and learn how to identify PHI from the data itself rather than relying on precompiled, manually-constructed sets of data. These latter systems may often serve a pre-processing role for de-identification, relying on subsequent manual processes to complete the de-identification steps missed by the automated system.(34) Annotated training data typically need to be produced specific to the particular type of EHR data to be de-identified, but software tools can enhance the process.(45) For both images and biological samples, there are too few studies with a paucity of quantitative data to judge the best approach, and the latter data category has the added Common Rule anonymization requirements needed for IRB exemption that do not appear to be satisfactorily addressed by the current approaches. Lastly, these strategies will continue to improve since the HHS is in the process of revising the rules protecting research subjects and the HHS Office for Civil Rights has held recent conferences on de-identification standards and HIPAA privacy.

How essential is anonymization?

In theory, anonymization is important since it places the patient's or research participant's right to privacy as the top priority in any anticipated or unanticipated scenario, and dramatically minimizes the release of sensitive information that may discriminate or stigmatize the individual from a social or economic perspective. In practice, it still may be possible to identify an individual from supposedly anonymized data sets, especially with respect to rare diseases within a specific geographical area. Additionally, the inability to re-identify individuals with their data may hamper the ability of investigators to conduct further studies on given individuals and pose bioethical challenges regarding the inability of clinicians to inform patients of results uncovered in their studies that may be relevant to their future health or well-being. Strategies which use pseudonymization rather than true anonymization may help to resolve these constraints, since they allow the data to be associated with a patient only under specified and controlled circumstances.

Do de-identification strategies alone meet the needs of multicenter research studies, including comparative effectiveness research?

This question has two components:

- *Besides the de-identification of individual documents, what can be done to ensure the privacy of data sets?* Heuristic methods for data de-identification are frequently used to comply with the HIPAA Privacy Rule; these include creation of a de-identified data set or limited data set, or generation of variables to replace identifiers (HIPAA de-identification relinking fields must not use hash codes, though limited data sets may) as long as the codes prevent identification of individuals by the data recipient). There are approaches that use one or more of these methods, including, for example, systems have been developed in which the user can request specific fields and records and a database is generated with information matching the anonymity level set by the user with respect to the recipient profile. Additionally, it has been argued by Ferris and others that to protect PHI and to construct a useful clinical research database, a hybrid system approach utilizing secure key escrow, de-identification, and role-based access for institutional review board (IRB)-approached researchers may be preferable in order to offer flexible control of PHI while meeting the needs of biomedical researchers.

- *What approaches can be used on a multicenter level to ensure patient or participant privacy?* De-identification and anonymization strategies comprise one important component of an integrated data collection and management system used in multicenter research studies and CER. In addition, some institutions use honest brokers, which collect and provide data to research investigators in a manner whereby it would not be reasonably possible for investigators to identify the participants directly or indirectly. Honest brokers are not part of either the clinical or research team since the honest broker links research identifiers and clinical identifiers. Honest brokers or coordinating centers of multicenter studies may also use universal identifiers or a master patient index to associate various patient identifiers across disparate systems, thus enabling the exchange of data between participating EHR systems and centers.

What is an important area for further work?

One critical area is the management of identifiers for the protection of genetic information, particularly with respect to protecting the privacy of identities to which DNA sequences were derived. This area of genomic privacy is particularly challenging for the biomedical community, given the immense quantity of data that needs to be processed, stored, and shared, as well as the consequences that identifying genomic data may have on an individual's health, employment, and insurance status.

Acknowledgments

Funding was provided by a contract from AcademyHealth. Additional funding was provided by AHRQ 1R01HS1973 (Comparative Outcomes Management with Electronic Data Technology (COMET) Study).

REFERENCES

1. Sweeney L Computational disclosure control: A primer on data privacy protection. Massachusetts Institute of Technology; 2001
2. Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32 [PubMed: 18652655]
3. Velupillai S, Dalianis H, Hassel M, et al. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. Int J Med Inform 2009;78:e19–26 [PubMed: 19482543]
4. Dalianis H, Velupillai S. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. J Biomed Semantics 2010;1:6 [PubMed: 20618985]
5. Grouin C, Rosier A, Dameron O, et al. Testing tactics to localize de-identification. Stud Health Technol Inform 2009;150:735–739 [PubMed: 19745408]
6. Tu K, Klein-Geltink J, Mitiku TF, et al. De-identification of primary care electronic medical records free-text data in Ontario, Canada. BMC Med Inform Decis Mak 2010;10:35 [PubMed: 20565894]
7. Miller R, Boitnott JK, Moore GW. Web-based free-text query system for surgical pathology reports with automatic case deidentification. Arch Pathol Lab Med (abstract) 2001;125:1011
8. Thomas SM, Mamlin B, Schadow G, et al. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002:777–781 [PubMed: 12463930]
9. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127:680–686 [PubMed: 12741890]
10. Douglass M, Clifford GD, Reisner A, et al. Computer-assisted de-identification of free text in the MIMIC II database In: Murray A, ed. Computers in Cardiology. Chicago, IL; 2004:341–344
11. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121:176–186 [PubMed: 14983930]
12. Douglass MM, Cliffford GD, Reisner A, et al. De-identification algorithm for free-text nursing notes Computers in Cardiology. Lyon, France; 2005:331–334
13. Sweeney JP, Portell KS, Houck JA, et al. Patient note deidentification using a find-and-replace iterative process. J Healthc Inf Manag 2005;19:65–70 [PubMed: 16045086]

14. Beckwith BA, Mahaadevan R, Balis UJ, et al. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12 [PubMed: 16515714]

15. Huang LC, Chu HC, Lien CY, et al. Embedding a hiding function in a portable electronic health record for privacy preservation. J Med Syst 2010;34:313–320 [PubMed: 20503616]

16. Dorr DA, Phillips WF, Phansalkar S, et al. Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf Med 2006;45:246–252 [PubMed: 16685332]

17. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84:362–369 [PubMed: 18500243]

18. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc 2008;15:601–610 [PubMed: 18579831]

19. Fielstein EM, Brown SH, Speroff T. Algorithmic de-identification of VA medical exam text for HIPAA privacy compliance: Preliminary findings In: Fieschi M, ed. MEDINFO 2004 Amsterdam: IOS Press; 2004:1590

20. Sibanda TC. Was the patient cured? Understanding semantic categories and their relationships in patient records. Electrical Engineering and Computer Science. Boston, MA: Massachusetts Institute of Technology; 2006:107

21. Sweeney L Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996:333–337 [PubMed: 8947683]

22. Ruch P, Baud RH, Rassinoux AM, et al. Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000:729–733 [PubMed: 11079980]

23. Morrison FP, Li L, Lai AM, et al. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc 2009;16:37–39 [PubMed: 18952938]

24. Morrison FP, Sengupta S, Hripcsak G. Using a pipeline to improve de-identification performance. AMIA Annu Symp Proc 2009;2009:447–451 [PubMed: 20351897]

25. Pestian JP, Itert L, Andersen C, et al. Preparing clinical text for use in biomedical research. J of Database Management 2006;17:1–11

26. Uzuner O, Sibanda TC, Luo Y, et al. A de-identifier for medical discharge summaries. Artif Intell Med 2008;42:13–35 [PubMed: 18053696]

27. Sleator D, Temperley D. Parsing english with a link grammar Technical Report CMU-CS-91–196. Pittsburgh, PA: Computer Science Department, Carnegie Mellon University; 1991

28. Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp 2002:757–761 [PubMed: 12463926]

29. Sibanda T, He T, Szolovits P, et al. Syntactically-informed semantic category recognition in discharge summaries. AMIA Annu Symp Proc 2006:714–718 [PubMed: 17238434]

30. Gardner J, Xiong L. HIDE: An integrated system for Health Information DE-identification Computer-Based Medical Systems, 2008 CBMS '08 21st IEEE International Symposium on Computer-Based Medical Systems. Jyvaskyla, Finland; 2008

31. Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010;79:849–859 [PubMed: 20951082]

32. Yeniterzi R, Aberdeen J, Bayer S, et al. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc 2010;17:159–168 [PubMed: 20190058]

33. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14:550–563 [PubMed: 17600094]

34. Wellner B, Huyck M, Mardis S, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007;14:564–573 [PubMed: 17600096]

35. Engelstad ME, McClellan M, Jacko JA, et al. Deidentification of Facial Images Using Composites. J Oral Maxillofac Surg 2011

36. Prior FW, Brunsden B, Hildebolt C, et al. Facial recognition from volume-rendered magnetic resonance imaging data. IEEE Trans Inf Technol Biomed 2009;13:5–9 [PubMed: 19129018]

37. Bischoff-Grethe A, Ozyurt IB, Busa E, et al. A technique for the deidentification of structural brain MR images. Hum Brain Mapp 2007;28:892–903 [PubMed: 17295313]

38. Clark KW, Gierada DS, Marquez G, et al. Collecting 48,000 CT exams for the lung screening study of the National Lung Screening Trial. J Digit Imaging 2009;22:667–680 [PubMed: 18777192]

39. Bland PH, Laderach GE, Meyer CR. A web-based interface for communication of data between the clinical and research environments without revealing identifying information. Acad Radiol 2007;14:757–764 [PubMed: 17502264]

40. Gonzalez DR, Carpenter T, van Hemert JI, et al. An open source toolkit for medical imaging de-identification. Eur Radiol 2010;20:1896–1904 [PubMed: 20204640]

41. Zhu Y, Singh PD, Siddiqui K, et al. An automatic system to detect and extract text in medical images for de-identification In: Liu BJ, Boonn WW, eds. Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications. San Diego, CA; 2010

42. Gulcher JR, Kristjansson K, Gudbjartsson H, et al. Protection of privacy by third-party encryption in genetic research in Iceland. Eur J Hum Genet 2000;8:739–742 [PubMed: 11039572]

43. Furuta K, Yokozawa K, Takada T, et al. De-identification procedure and sample quality of the post-clinical test samples at the bio-repository of the National Cancer Center Hospital (NCCH) in Tokyo. Jpn J Clin Oncol 2011;41:295–298 [PubMed: 20852301]

44. Hara K, Ohe K, Kadowaki T, et al. Establishment of a method of anonymization of DNA samples in genetic research. J Hum Genet 2003;48:327–330 [PubMed: 12750962]

45. Mayer J, Shen S, South BR, et al. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes. AMIA Annu Symp Proc 2009;2009:416–420 [PubMed: 20351891]

46. Aramaki E, Imai T, Miyo K, et al. Automatic deidentification by using sentence features and label consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data 2006

47. Guo Y, Gaizauskas R, Roberts I, et al. Identifying personal health information using support vector machines. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data 2006

48. Hara K. Applying a SVM based chunker and a text classifier to the Deid Challenge; i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data; 2006.

49. 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]

50. Onken M, Riesmeier J, Engel M, et al. Reversible anonymization of DICOM images using automatically generated policies. Stud Health Technol Inform 2009;150:861–865 [PubMed: 19745435]

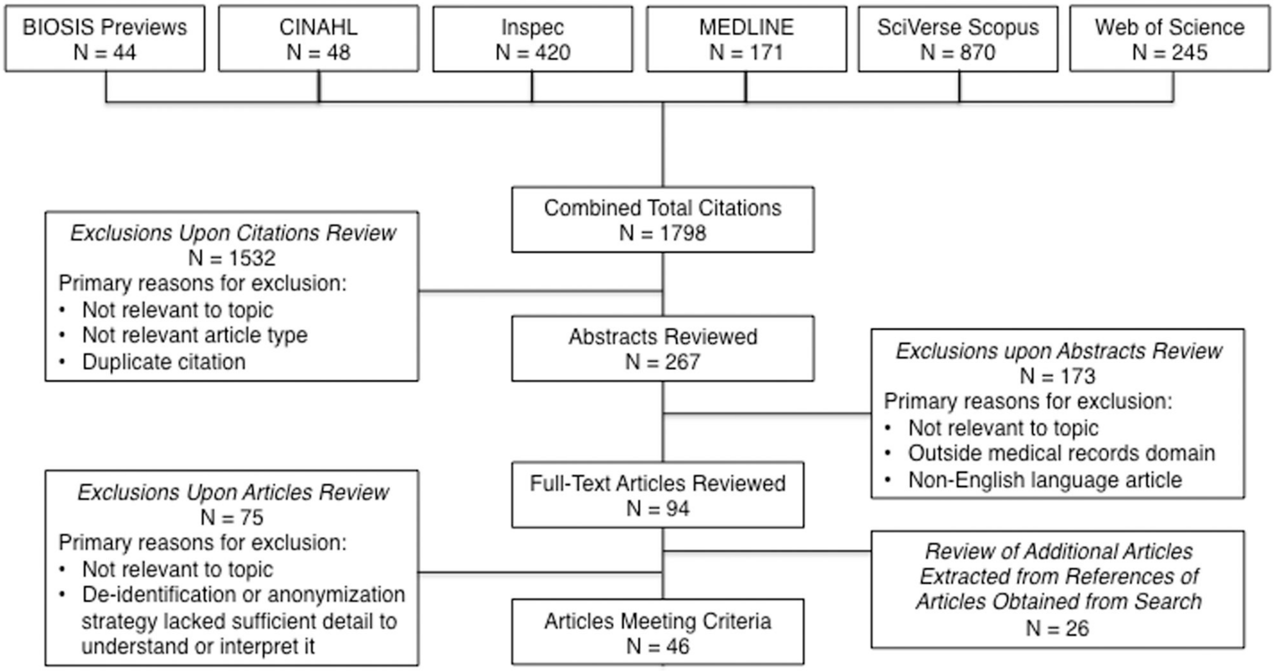
BIOSIS Previews
N = 44
CINAHL
N = 48
Inspec
N = 420
MEDLINE
N = 171
SciVerse Scopus
N = 870
Web of Science
N = 245
Combined Total Citations
N = 1798
Exclusions Upon Citations Review
N = 1532
Primary reasons for exclusion:
• Not relevant to topic
• Not relevant article type
• Duplicate citation
Abstracts Reviewed
N = 267
Exclusions upon Abstracts Review
N = 173
Primary reasons for exclusion:
• Not relevant to topic
• Outside medical records domain
• Non-English language article
Full-Text Articles Reviewed
N = 94
Exclusions Upon Articles Review
N = 75
Primary reasons for exclusion:
• Not relevant to topic
• De-identification or anonymization strategy lacked sufficient detail to understand or interpret it
Review of Additional Articles Extracted from References of Articles Obtained from Search
N = 26
Articles Meeting Criteria
N = 46

Figure 1.

Flow Diagram of Search Results

Table 1.

Definitions of Various Data Protection Methods

Data Protection	Description
Anonymization	Irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link
Augmentation	Often achieved by generalization, in which each record is indistinguishable from another shared record
Binning	Data pre-processing technique used to reduce the effects of minor observational errors; the original data values which fall in a given small interval (i.e., a bin) are replaced by a value representative of that interval
Cell Suppression	Blanking certain fields in a data table in such a way that no entry (row) in the table is unique
Censoring	Value of a measurement or observation is only partially known
De-identification	Removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data
Depersonalization	Process of identifying and separating personal from other data
Disambiguation	Process to provide clarity when a term is ambiguous
Encryption	Process of transforming data using an algorithm to make it unreadable except to the intended recipient
Eponyms	Words that could be both clinical terms used in a report or proper names (e.g., Parkinson’s disease)
Exclusion	Prohibition of specific data elements
Generalization	Process of creating successive layers of summary data in a database
Hash Function	Algorithm that converts a large data set into a small datum, usually a single integer that may serve as an index to an array; typically resulting in an anonymous code, same for a given individual, but impossible to retrieve the identity
Hiding Function	Methodology that makes information invisible except to the intended recipient
Obfuscation	Concealment of intended meaning in communication, making communication confusing, intentionally ambiguous, and more difficult to interpret
Pseudonymization	Identification data is transformed and then replaced by a specifier that cannot be associated with the data without knowing a certain key
Transformation	Conversion of data from a source data format into destination data

Table 2.

Summary Table*

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
Sweeney, Massachusetts Institute of Technology, Cambridge, MA (1996) (21)	Text	Pediatric medical records including letters to referring physicians	The Scrub System uses numerous detection algorithms with local knowledge competing in parallel to label contiguous text characters as being identifiers; each detection algorithm recognizes a specific entity, where recognizable entities are fields such as first name, last name, street address, and date, and replaces PHI with a pseudo-value replacement text	Scrub system detected 99-100% of all PHI compared to a straightforward approach of global search and replace that detected no more than 30-60% of all PHI in a scrubbed subset of 275 patient records and included 3,198 letters to referring physicians	Limited testing; test corpus details and false positive not reported; require time/staff/local knowledge for creation of lists for detection algorithms
Ruch, University Hospital of Geneva, Switzerland (2000) (22)	Text	Post-operative reports, laboratory and test results, and discharge summaries mainly in French, but also in English (<1%) and German (<1%)	Natural language Processing (NLP) tools provided by the MEDTAG lexicon (based on the UMLS Metathesaurus) that includes a semantic lexicon specialized in medicine and a toolkit for word-sense and morpho-syntactic tagging of words for disambiguation task; disambiguation rules assume that syntax can help to distinguish meanings of words with different syntactic categories, simple taggers can help solve syntactic and semantic ambiguities, and information extraction can also be tag-assisted; extraction and removal of confidential items uses formal recursive transition networks (RTN)	System tested on data that included 467 identifiers in approximately 800 documents: 96.8% identifiers correctly removed, 1.7% identifiers removed with removing also irrelevant tokens, 0.6% identifiers incompletely removed, 0.9% identifiers left in the text, 0% tokens removed, which are not identifiers; system designed to work in a multilingual environment	Limited testing on data set; tractability of anonymization not allowed by system, which is a security advantage, but authors comment that tractability may be necessary for studies on genealogy; 40 disambiguation rules written manually took 3 weeks to write
Miller, Johns Hopkins Medical Institutions, Baltimore, MD (2001) (7)	Text	Surgical pathology reports	System scrubs proper names in a free-text database of surgical pathology reports; names identified by lists of persons, places, and institutions, or by proximity to key words are replaced by suitable tokens prior to display on the web-based query system	6.6% of 361,957 surgical pathology cases contained proper names that were tokenized by the de-identification system as of June 1, 2000	Very unusual combination of diseases may identify participant; all names and all misspellings must be captured and included in the dictionary
Taira, University of California, Los Angeles, CA (2002) (27)	Text	Pediatric urology reports: letters and reports to referring physicians, discharge summaries, clinical notes, and operative/surgical reports	Algorithm (using NLP tools) estimates the fitness of candidate patient name references by establishing a set of semantic selectional restrictions that place tight contextual requirements upon candidate words in the report text (1,350 manually tagged training reports); Using a maximum entropy probabilistic model, probabilities are assigned based on how well a given candidate satisfies the set of semantic restrictions; refers to a lexicon of over 64,000 first and last names;	Of 900 random test reports, patient name references for each report were hand-tagged (gold standard) and compared to the system: area under ROC curve is 0.9735; best overall performance at a decision threshold of 0.55 at which recall was 93.9% and precision was 99.2%; time to process a 5Kb report was 30 sec on a 2GHz computer	Testing limited to patient names in the domain of pediatric urology; identification of erroneous logical relation instances can possibly propagate false positive errors; required manual tagging of 1350 training reports

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
Thomas, Regenstrief Institute for Health Care, Indianapolis, IN (2002) (8)	Text	Pathology and Cytology reports transformed into XML documents	Search and replace algorithm designed using a substitution method built on the assumption that proper names almost always occur in pairs in clinical reports; words in Clinical and Common Usage Words (CCUW) list (built from UMLS word index and word list from the GNU spellchecking program Ispell; 320,000 words) retained in reports unless surrounding word is a proper name, and words in list of proper names (from Ispell, Regenstrief Institute Medical Record System, Social Security Death Index; 1.8 million names) were removed from reports	Upon searching the entire report (108,092 words), the algorithm correctly identified 92.7% of 7,710 proper names, missed 7.3% of proper names, and incorrectly marked 1.9% as proper names	Removes only proper names; report characteristics at home institution make the system reliable, but this approach may not work at other institutions; large amount of words (37,000) common to both CCUW and proper name lists had to be addressed; large number of false positives
Berman, National Cancer Institute, Rockville, MD (2003) (9)	Text	Surgical pathology reports	The Concept-Match Algorithm (written in Perl) replaces a nomenclature code and a synonym for a medical term matching a standard nomenclature term (found in UMLS; could also use algorithm with MESH or SNOMED); high-frequency “stop” words (e.g., a, an, the, for) are left in place; any other words are replaced by asterisks	Open-source implementation of the algorithm is freely available; scrubbed 567,921 surgical pathology report phrases in less than 1 hour	Limitations common to many text scrubbers that have potential solutions: the meaning of a sentence is almost always changed and sometimes lost when words are removed or altered; autocoding is not perfect, thus miscoding is inevitable; reference nomenclature containing identifying or incriminating concepts may appear in the scrubbed output
Douglass, Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA (2004) (10)	Text	Nursing notes from intensive care unit	Perl-written algorithm uses pattern-matching to identify potential dates, telephone numbers, SS#, and other protected types of identification numbers; uses look-up tables to identify potential locations and patient, clinician, and hospital names; applies several simple context-based rules; replaces PHI type with an appropriate but randomly chosen text replacement	Open source; Corpus was 2,646 nursing notes from 148 randomly selected patients; recall and precision, respectively, of de-identification strategies manually by clinicians vs. algorithm: single clinician 81% and 98%; 2 clinicians 94% and 97%; 3 clinicians 98% and 97%; algorithm 85% and 37%	Low precision rate since algorithm identified many false positives
Gupta, University of Pittsburgh Medical Center, PA (2004) (11)	Text	Surgical pathology reports	The De-Id system locates unstructured identifiable text, uses rules (including use of the UMLS Metathesaurus for identifying medical phrases) and dictionaries (including a user-customizable dictionary for local terms) to identify 17/18 HIPAA-specific identifiers, and replaces identifiable text with de-identified but specific tags	The de-identification engine had three evaluation rounds resulting in a successful tool for de-identifying reports with limited overmarking of important clinical information; De-Id has been used to de-identify > 35,000 pathology reports	Unable to identify/remove full-face photographic images; unable to detect typographical errors; quality assurance procedures not automated; no quantification of results
Fielstein, Department of Veterans Affairs, Tennessee Valley Healthcare System,	Text	VA compensation and pension joint exams transcribed to text	Perl-based algorithms using search-and-replace style pattern-matching techniques; utilized pattern-matching expressions from public domain sources (e.g. valid email addresses, census-based names, state names and abbreviations), and a local set of matching	Corpus was 69 VA exams sampled from a set of over 120,000 disability examinations extracted from 128 VA sites for quality improvement purposes; initial recall and specificity of the algorithms were 81.0% and	Detection of name and place misspellings were an issue; the authors report that adding phonetic and fuzzy matching techniques may be potential solutions

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
Nashville, TN (2004) (19)			patterns for VA-specific personal identifiers (e.g., case numbers)	99.9%, respectively; following revision to improve the recall of the algorithms for city names and dates, the overall recall and specificity were 92% and 99.9%, respectively	
Douglass, Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA (2005) (12)	Text	Nursing notes from intensive care unit	Improved version of previous system from 2004; Perl-written algorithm uses lexical look-up tables (e.g., 1990 U.S. Census, list of local hospitals, UMLS), regular expressions, and simple heuristics to locate multiple sources of PHI; displays all choices made by the algorithm and allows user to decide whether the identified PHI should be removed	Open source; corpus was 747 nursing notes from 22 patients; recall was 92% and precision was 44%; method to handle misspellings (e.g., using 100 most popular male and female first names)	High false positive rate was mainly due to repeated occurrences of incorrectly identified PHI and the lack of lists of common words and abbreviations found in nursing notes
Sweeney, Mayo Clinic, Jacksonville, FL (2005) (13)	Text	EHR in the Medical Data Trust, including emails, prescription refills, laboratory reports, and letters	SPANS (Simple PAtient Note Scrubber), written in Perl, uses a find-and-replace process with an iterative approach, in which each note is fed in as a single line and then multiple passes are performed; uses terms from text files in conjunction with the format of the surrounding words to identify text to be scrubbed; scrubs patient name, street and postal addresses, SS#, phone numbers, dates, age (if > 88 or < 18 years), occupation, job or company title, and email and website addresses	One test on a random sample resulted in 7 pieces of identifiable information in 400 notes, yielding an error rate of 1.75%; a second test resulted in 2 pieces of identifiable information, an error rate of 0.5%; using the higher error rate (1.75%) researchers were confident that the true error rate was between 0 and 6.75%, resulting in a statistical "safe harbor"; balance between preserving the anonymity of patients, while maintaining the context of the note; flexible system that may adapt to new terms and word sets	Identification of document type could tailor the logic to increase accuracy and efficiency; case sensitive searching in SPANS has difficulty with notes written in all capital letters; segregation of the different types of patient-specific data and tailoring the search logic for specific data types might produce a more accurate scrubbing process; conversion of large data set can take > 3 weeks to complete the process
Aramaki, University of Tokyo, Japan (2006) (45)	Text	Medical discharge summaries convert XML to IOB2 format	Manual tagging of all words of a training set with either a PHI tag or a non-PHI tag, followed by a machine learning CRF to learn the relation between features and labels in this tagged set; uses local (surrounding words and care), extra resource (dictionaries), and non-local (surrounding sentences position, length, and last sentence) features in addition to label consistency (preferring the same label for the same word sequence)	Corpus was 671 discharge summaries with 14,309 PHI tags; recall was 96.66%, precision was 98.29%, and F-measure 0.9747	Label consistency may not help differentiate PHI from non-PHI given the levels of ambiguity between PHI and non-PHI within individual EHRs; only small dictionaries were used in experiment- authors indicated they could achieve higher performance with larger dictionaries
Beckwith, Harvard University, Boston, MA (2006) (14)	Text	Surgical pathology reports converted to SPIN XML schema	HIPAA- compliant scrubber that searches/ removes occurrences of identifiers initially identified in the report header, searches for predictable patterns likely to represent identifying data, and compares each word to a database of personal and geographic place names (>101,000 entries)	Open source under GNU GPL terms; 98.3% of 3499 unique identifiers in the test set of 1800 reports were removed, with only 19 HIPAA-specified identifiers missed; scrubbing speed of 47 cases per minute; over-scrubs of 2.6 incorrectly removed phrases per report	Unable to remove misspelled names and difficult to remove accession numbers (with diverse formats) and foreign addresses; new terms (e.g., Her-2) may be mistaken for identifiers; medical eponyms (e.g., Brown) are challenging for all scrubbers; quality control measure would be highly desirable

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
Dorr, Oregon Health and Science University, Portland, OR (2006) (16)	Text	Inpatient, outpatient, and administrative notes	Manual and simple automated de-identification techniques were compared: 2 independent reviewers used simple search and replace algorithms and visual scanning to find HIPAA-defined PHI, followed by an independent second review to detect any missed PHI vs. a simple automated tagging program using regular expressions was run for number- or date-based PHI	Corpus was 262 notes randomly selected from a research database containing 88,000 distinct patient notes; for manual review: recall of 95.9%, specificity of 99.99%, and average precision of 99.6% with 7/1996 incorrect PHI assignments, and excellent inter-rater agreement for PHI number and type (ICC 0.99, 95% CI 0.98-1.00); for automated system: recall of 98.5%, specificity of 99.8%, and average precision of 88.4% with 168/1442 incorrect PHI assignments	Manually de-identification difficult and time consuming: manually de-identifying a note took 87.3±61 seconds, and, by extrapolation, de-identifying entire dataset (88,000 notes) would take 58 weeks+1 day, assuming 7.5 hours/day × a 5-day work week; simple automated techniques did not perform as well as human reviewers
Guo, Sheffield University, Sheffield, UK (2006) (46)	Text	Medical discharge summaries	Uses GATE system, a NLP framework and model in addition to an information extraction system to preprocess and annotate a training set; an SVM classifier and a set of rules that incorporate entity-specific knowledge for mapping the named entities recognized by the information extraction system to PHI categories were also utilized	Open source; for two outputs (the first run used only token level features, the second run also incorporated additional features), the weighted F-measure was 0.987 and 0.982	The authors report that the system sometimes performs less well when trained with all the available training data rather than some portion of it, hinting at a problem of over-fitting
Hara, Nara Institute of Science and Technology, Nara, Japan (2006) (47)	Text	Medical discharge summaries	Four processes are employed: determination of headings by pattern matching, string pattern matching using regular expressions, sentence classification by syntactic structures that classifies sentences as containing PHI or not, and phrase chunking (or sequence labeling) that identifies PHI categories based on SVM	Corpus was 891 discharge summaries; for three outputs, the best run achieved F-measures of ≥ 0.90 for 8 PHI categories, except for phone (0.867), hospital (0.882), and location (0.677)	Sentence classification may impair performance
Pestian, University of Cincinnati, OH (2006) (25)	Text	Clinical notes from pediatric patient encounters	Encryption Broker (EB) software, containing a data-cleansing algorithm, including disambiguating the unstructured clinical text containing jargon and acronyms by a vast number of rules analyzing surrounding words, followed by data scrubbing, in which 16 PHI elements were eliminated or replaced using systemic bias as a method of encryption (e.g., all female names changed to “Jane”)	Open source software; EB was evaluated by 348 randomly selected records paired with the corresponding data output from EB; 10,240 paired sentences were reviewed by clinical experts, and EB correctly changed a token 98% of the time, and when a token should not be changed, it was not 99% of the time	EB relies on manual rule development; the authors indicate that the more than 1,000 disambiguation rules had to be developed as the first set; the rules may not be generalizable to other locations or patient population; the majority of errors were related to ambiguous names which are problematic for all systems
Sibanda, Massachusetts Institute of Technology, Cambridge, MA (2006a) (20)	Text	Medical discharge summaries	Statistical de-identifier for medical discharge summaries using local lexical and syntactic context; a multi-class SVM was trained on human-annotated data to classify each word as PHI vs. non-PHI (a binary classification task)	For authentic discharge summaries: F-measure of 0.9682, recall of 95.24%, precision of 98.31% on 4,243 instances of PHI; F-measure of 0.9990, recall of 99.95%, precision of 99.84% on 112,669 instances of non-PHI	Limited to 7 classes: patients (patient first and last names, health proxies, family members), doctors (MDs and other practitioners), hospital and specific medical organization names, IDs (number-letter combinations identifying medical

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
					records, patients, doctors, hospitals), dates, location, phone numbers
Sibanda, Massachusetts Institute of Technology, Cambridge, MA (2006b) (28)	Text	Medical discharge summaries	Semantic category recognition approach trained with syntactic and lexical contextual clues and UMLS ontological information for document understanding that analyzes the syntax of documents; statistical semantic category recognizer identifies 8 semantic categories (diseases; signs, and symptoms; tests; treatments; results; dosage information; abusive substances; and medical practitioners) in the discharge summaries; individual words classified as one of the 8 semantic categories or as the none category	48 summaries with 5,166 sentences; for the 8 categories, the classification F-measure range was 0.832 to 0.956, with a recall range of 80.9% to 96.2% and a precision range of 85.7% to 97.8%; for clinical text, contextual clues (lexical and syntactic) provide stronger indications of semantic categories than information extracted from UMLS	Words occurring infrequently were misclassified, which the authors attribute to the susceptibility of statistical approaches to paucity of data; the results and signs and symptoms categories still pose problems
Szarvas, University of Szeged, Hungary (2007) (48)	Text	Discharge records	A machine learning-based iterative NER approach intended for use on semi-structured documents like discharge records; labels all entities whose tags can be inferred from the structure of the text (based on contextual and surface patterns without reference to deep knowledge information) and it then utilizes this information to feed a decision tree learning algorithm to find further PHI phrases in the flow text parts of the document; includes regular expressions for the well-defined classes and subject heading information	For 9-way evaluation measuring performance of identifying 8 PHI classes and non-PHI class as well, an F measure score of 99.4814% was attained; did not use deep knowledge information (e.g., POS codes) or domain specific resources	Significantly better on precision than recall regarding the recognition of the 8 PHI classes; did not recognize approximately 0.18% of PHI; typical errors were misclassifications (e.g., doctor names that were common words), confusion between classes (location vs. hospital), or the inability of the statistical modes to handle ambiguous text based on contextual patterns
Wellner, MITRE Corporation, Bedford, MA (2007) (33)	Text	Medical discharge summaries	Two NER toolkits were evaluated: Carafe from MITRE implements CRFs targeted especially to phrase identification tasks, and LingPipe from Alias-I performs a variety of NLP tasks and uses named-entity tagging by chunkers that operate on n-gram based character language models (hidden Markov model + statistics from training)	Corpus were 910 medical discharge summaries (690 used for training, 220 used for testing); the “out of the box” Carafe system achieved a phrase F-measure of 0.9664 with only 4 hours of work to adapt it to the task; further work reduced the token level error term by over 36%, achieving a phrase F-measure of 0.9736.	Although the total number of errors produced by Carafe was small, 50% of the errors were missing tags, that can result in a disclosure of PHI
Friedlin, Regenstrief Institute, Indianapolis, IN (2008) (18)	Text	HL7 messages from clinical documents including narrative text documents	The MeDS (Medical De-identification System) scrubs reports through a series of scrubbing processes involving pre-processing (extraction of patient data from patient information section, text normalization), pattern matching algorithms (to detect SSN, address, dates), name matching to proper name lists vs. CCUW lists (to identify patient, provider), report-specific matching (detection of specific patient data extracted from information section), and text string nearness	In addition to HIPAA-specified identifiers, also removes health care provider names, personnel identifiers, institution names, and all references to ages and times; initially scrubbed 99.06% of 11,380 HIPAA-specified identifiers and 98.26% of 38,768 of non-HIPAA-specified identifiers; after software modification, scrubbed 99.47% of 80,418 HIPAA-specified identifiers and 96.93% of 13,091 non-	Unable to remove biometric identifiers and full-face photographs (not contained in text-only messages); reports were regional (central Indiana area); developer of software acted as gold standard and scrubbing process evaluator; name nearness scrubber has many false positives

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
			algorithms for names (to identify spelling and typographical errors)	HIPAA-specified identifiers; over-scrubbing rate 7% of 54,160 identifiers	
Gardner, Emory University, Atlanta, GA (2008) (29)	Text	Pathology reports	The HIDE (Health Information DE-identification)) System can be used on both structured and unstructured data via data linking (person-centric identifier view with identifying attributes linked to each individual), identifying and sensitive information extraction (statistical learning approach using a CRF-based NER for unstructured data), and anonymization (data removal and generalization)	Corpus of 100 textual pathology reports; 98.2% of 2,725 attributes extracted, including name, age, date, account number, and medical record number	Proof-of-concept study using limited amount of data; still a challenge to extract indirectly identifying information
Neamatullah, Massachusetts Institute of Technology, Cambridge, MA (2008) (2)	Text	Nursing notes from intensive care unit	The Perl-based deid software package performs a lexical match on each word in the text with dictionaries containing known/potential PHI look-up tables to locate PHI (labeled with the associated dictionary type); pattern matching is performed using regular expressions and heuristics that search for patterns with various contextual keywords to find more named entities; each instance of PHI is replaced with a tag to indicated its corresponding category	Open source license; performance on development corpus of 2,434 nursing notes containing 334,000 words and 1,779 instances of PHI yielded recall of 96.7%, precision of 74.9%, and fallout of 0.2%; performance on test corpus of 1,836 nursing notes containing 296,400 words yielded recall of 94.3% and 78 instances of false negatives	Authors report that although the algorithm accuracy is high, it is probably insufficient to be used to publicly disseminate medical data; algorithm depends on customized local dictionaries; misspellings continue to be a challenge in de-identifying all free text
Uzuner, University at Albany, State University of New York, Albany, NY (2008) (49)	Text	Medical discharge summaries	The Stat De-id system uses support vector machines to classify each word in a sentence as belonging to one of 8 categories (doctor, location, phone, date, patient, ID, hospital, non-PHI); contextual clues human annotators found useful in de-identification are captured by target words and features of the words surrounding the target; syntactic information is extracted from the Link Grammar Parser, and is augmented with medical dictionaries	Five corpora were used, three were developed from a corpus of 48 discharge summaries, the fourth was from 90 discharge summaries of deceased patients, and the fifth from 889 de-identified discharge summaries; Stat De-id recognizes 94-98% of PHI	Stat De-id performs relatively poorly in recognizing phone and location PHI classes
Grouin, Centre National de la Recherche Scientifique, Orsay, France (2009) (5)	Text	Cardiology reports (in French)	Medina (Medical Information Anonymization) is based on deid (Neamatullah, 2008) (2) and MeDS (Friedlin, 2008)(18) systems; regular expressions, list, and dictionaries were created and words that occur in the common words dictionary (e.g., English last names that do not exist in French) were removed from the lists	Corpus for evaluation was 23 randomly selected texts from 21,749 clinical texts; recall and precision, respectively, for first-level de-identification and removal of patient name and birth date (91%, 100%), Medina (83%, 92%), and a French version of the de-id system (65%, 23%)	Small test sample; deid system was unable to be well-adapted from English to French since investigators were unable to efficiently modify regular expressions for French, resulting in over-de-identification; authors suggestion Medina recall can be improved by adding French resources
Morrison, Columbia University, New	Text	Outpatient clinical follow-up notes and letters from internal medicine, infectious disease,	MedLEE Processor, an existing general purpose NLP system, was used without modification to process a heterogeneous set of	Of 8 PHI types (patient name, clinician name, hospital, identifiers [e.g., SS#], date [except year],	Required preprocessing, which involved approximately 10-15 hours of programming to extract

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
York, NY (2009a) (23)		oncology, hematology, neurology, cardiology	outpatient clinical notes into XML-tagged clinical data	location, phone number, and age > 89 years), there were 809 instances of PHI in the 100 outpatient notes, 26 instances (3.2%) of PHI was allowed into the MedLEE output	the text from the database and remove unrecognizable characters; the authors reports that MedLEE output is not necessarily de-identified and should not be treated as such, particularly since the rate of PHI that was allowed into output using MedLEE is higher than other systems; pairing two systems with different strategies may produce better results
Morrison, Columbia University, New York, NY (2009b) (24)	Text	Outpatient clinical follow-up notes and letters from internal medicine, infectious disease, oncology, hematology, neurology, cardiology	MedLEE Processor used in a prior study (Morrison, 2009a)(23) was paired with deid (Neamatullah, 2008)(2), so that MedLEE was run on the unmodified text output from deid, resulting in XML-tagged parsed concepts	For a total of 100 notes with 818 PHI instances, deid alone missed 24% of PHI in text, resulting in 7% total misses of patients'/clinicians' names; with both deid and MedLEE used serially, only 2.1% of PHI was missed in text, and the error rate for names dropped from 3.5% with MedLEE alone to 0.7% with both systems	Study did not use local dictionaries to optimize deid performance; age and pager numbers were not handled well by the systems, nor were names that are common English words and not in a context that is easy to identify as a name
Velupillai, Stockholm University, Sweden (2009) (3)	Text	EHR from neurology, orthopedics, infection, dental surgery, and nutrition clinics	Deid-Swe is a modified deid system (Neamatullah, 2008)(2) for EHR in Swedish; adapted system to manage Swedish phone numbers, SS#, and date formats; added Swedish lists	Corpus was 100 mainly free-text EHRs; fairly high Inter-Annotator Agreement (IAA) results on a manually-created gold standard, especially for specific tags such as names; the average IAA over all tags was 0.65 F-measure (0.84 F-measure highest pairwise agreement); for name tags the average IAA was 0.80 F-measure (0.91 F-measure highest pairwise agreement)	The de-identification software written for American English to Swedish directly yielded F-measures of 0.04-0.16, with precision of 0.03-0.09 and recall was 0.56-0.76
Aberdeen, Vanderbilt Unversity, Nashville, TN (2010) (30)	Text	Discharge summaries, laboratory reports, letters, and order summaries	MIST (MITRE Identification Scrubber Toolkit) uses examples of de-identified text (training data), from which the tool can learn optimal context-dependent features to ensure accurate identification of the entities to be redacted; the training exemplars are derived from annotated medical records	Open source; F-measure of 0.960, recall of 97.8%, precision of 94.3%, error rate of 1.0%; less amount of overhead and upkeep than other approaches, since it involves only adding new training documents as document formats evolve, and the subsequent retraining of the de-identification model; can be rapidly adapted to new domains and tasks	Some knowledge may be more efficiently captured from experts via use of rules, rather than by machine learning; the corpus evaluated was not a true gold standard and it is possible that errors in the data generated by this process could bias the evaluation
Huang, National Yang-Ming University, Taipei, Taiwan (2010) (15)	Text	Portable electronic health record, specifically discharge summaries	Several regular expression rules were used to perform pattern-matching to remove numerical identifiers; expression rules with specific characters and symbols were used to detect the identifier patterns combined with vehicle number, SS#, medical record, etc.;	182 admission/discharge summaries in free form type containing 158,303 words for recognizing identifiers were evaluated; 3,573 words of HIPAA identifiers were found; 57 regular expression rules and 68,328 keywords	Authorization requirements to identify legal users can be developed, but do not protect the data from being viewed by others while the user reads it; although the algorithm correctly recognizes

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
			keyword filter was used to match patterns to recognize the identifier and filter the context of matching keyword; after finding identifiers, JavaScript that is embedded into the portable EHR with the purpose of hiding the information was generated, and this allows the user to switch between displaying and hiding the identifiers	were defined; average number of identifiers was 19 in a record; averaged recall, precision and F-measure were 97.5%, 98.9%, and 0.98, respectively	identifiable information most of the time, it makes undermarking and overmarking errors, and failed when there was a typographic or spelling error
Dalianis, Stockholm University Forum, Sweden (2010) (4)	Text	EHR from neurology, orthopedics, infection, dental surgery, and nutrition clinics	Two gold standards of a manually annotated gold standard (Vellupillai, 2009) for de-identification, one created automatically, and one created through discussions among annotators; the gold standards were used for the training and evaluation of an automatic system (Stanford Named Entity Recognizer) based on the CRF algorithm	For both gold standards, an F-score around 0.80 was obtained for several experiments evaluating sets of around 4-6,000 annotation instances with four-fold cross-validation; 49 false positives that were verified true positives were found by the system but missed by the annotators	The default settings of the Stanford CRF was used for all experiments; the authors indicate that further analysis on and evaluation of useful and extended features as well as weighting schemes for this specific classification problem is needed
Tu, Institute for Clinical Evaluative Sciences, Toronto, Canada (2010) (6)	Text	Primary care records, including point-form progress notes, diagnostic tests, operative reports, consultation letters, and discharge summaries	Modified deid system (Neamatullah, 2008)(2) for an Ontario context for use on primary care EMR data by removing functionality to preserve dates, months written in text format changed to number format, replaced or added lists based on American context with Ontario lists, added 600 more medical eponyms, added code to prevent removal letters in acronyms or other medical terms, added a “do not remove” list	Modified program on a training set of 1,000 free-text records performed with a recall of 88.3%, specificity of 91.4%, precision of 91.3%, accuracy of 89.9% and F-measure of 0.90; two validation sets of 500 and 700 notes had sensitivities of 86.7% and 80.2%, specificities of 91.4% and 87.7%, precisions of 91.1% and 87.4%, accuracies of 89.0% and 83.8% and F-measures of 0.89 and 0.84, respectively	Searchable text would require additional processing if not in OCR format; addition of large lists to the program resulted in approximately 47 hours to process 5 years of data on 2,900 patients
Yeniterzi, Vanderbilt University, Nashville, TN (2011) (31)	Text	Discharge summaries, laboratory reports, letters, and order summaries	Further evaluation of MIST (MITRE Identification Scrubber Toolkit) (see Aberdeen, 2010)(30)	Open source; MIST was trained and tested on a collection of real and resynthesized Vanderbilt records; when trained/tested on real records, accuracy was 0.990 and F-measure was 0.960; when trained/tested on resynthesized records accuracy was 0.998 and F-measure was 0.980, indicating that MIST achieves high accuracy when training and test sets are homogeneous	MIST performance declined moderately when trained on real records and tested on resynthesized records with accuracy of 0.989 and F-measure of 0.862; results declined significantly when trained on resynthesized records and tested on real records with accuracy of 0.942 and F-measure of 0.728
Bischoff-Grethe, UC San Diego, CA (2007) (36)	Images	Brain MR images	Automated “defacing” algorithm that uses models of non-brain structures and removes only identifiable facial features from MR volumes	342 defaced datasets (from patients of various ages and diagnoses) had no brain tissue removed and upon inspection of the 3D images the identifying facial features (eyes, nose, mouth, chin) were removed; algorithm handles a variety of data formats and did not interfere with other automated techniques (e.g., skull-stripping);	Processing time may be greater than skull-stripping algorithms to de-identify MR images; currently can only be applied to T1-weighted datasets

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
				automated and scripted to process large datasets	
Bland, University of Michigan (2007) (38)	Images	DICOM data	Web-CAP is a custom-designed, de-identifying DICOM receiver that removes PHI, sets the value to null, or substitutes new, de-identified values for PHI, depending if the attribute containing the PHI is required or optional; patient's hospital ID is replaced with a research ID (unique identifier automatically assigned)	Software is implemented using standard web-based programming paradigms and conventional database technology, and is adaptable and customizable to meet institution-specific requirements	DICOM file de-identifier is a component used in the overall process of communicating PHI via Web-CAP; users could erroneously enter PHI in message boxes; no quantitative data provided
Onken, Institute for Information Technology, Oldenburg, Germany (2009) (50)	Images	DICOM data	DICOM images are "reversibly anonymized" by encryption and early removal of all identifying data, yet containing all relevant medical information; based on XML inputs, scripts were developed that automatically generated attribute lists for each DICOM object type, and a framework was implemented that read those files into an SQLite database and anonymized a given set of DICOM files; extracted identifying data and the anonymized DICOM object are marked by a token that allows for later reassembling of the original object	Authors report that it performs effectively and efficiently on real-world test images	No quantitative data provided
Clark, Washington University, Saint Louis, MO (2009) (37)	Images	Chest CT exams	Data displayed for any exam is obtained from its original DICOM header; when the user selects an exam to be de-identified, a Westat (Rockville, MD) Matching-List is searched until a line where DOB, gender, and exam date match those of the selected exam's DICOM header; if a match is made, then a new de-identified DICOM header is created, in which the participant ID replaces the medical center ID, the screening year is placed in a comment field, a generic string replaces the patient's name, a fixed string replaces the exam date, and the DOB, gender, accession number, and other DICOM fields containing PHI are blanked	The authors report that the vast majority of 48,547 exams were correctly de-identified of the DICOM elements	Exam dates in patient-protocol text-only image series, demographics in scout images, and radiology reports in secondary-capture image series were encountered; subsequently, the de-identification software was patched; however, no quantitative data provided
González, University of Edinburgh, UK (2010) (39)	Images	DICOM data (CT and MRI scans)	PrivacyGuard is a Java-programmed toolkit to de-identify DICOM data by removing personal information if not required, emptying their content, substituting the information for dummy values, or ambiguating their values; can also manage personal information in the pixel data contained in DICOM objects; open-source license; flexibility for different de-identification requirements	Successfully anonymized DICOM objects produced by different manufacturers' equipment, and successfully deployed in a distributed environment in which data were collected at 2 sites, stored locally, and then the data were anonymized and transferred to the processing site for storage	No quantitative data provided for the multicenter tests

Primary Author, Location (Year) (Citation)	**Type**	**Test Corpus, Other**	**Description**	**Strengths**	**Limitations**
Zhu, Syracuse University, Syracuse, NY (2010) (40)	Images	Ultrasound or secondary capture DICOM images	The region variances of intensity values for each pixel are calculated by the algorithm and a threshold is applied to detect text regions in the image, specifically to separate the text (PHI and annotations) from background and anatomic structures; in order to eliminate the over-detected text regions (i.e., non-text detected as text), post processing with heuristic rules is applied to the detected text after region grouping to remove mis-detected anatomic structures, lines, etc.; a level set method is used to extract PHI text from detected text regions for further OCR	128 images with PHI were tested, and the average recall rate is 99.05%; the average detection time for these images was 1.4456 seconds	Prototype only, which needs algorithm improvement, performance evaluation, computation optimization, and text recognition by OCR
Engelstad, University of Minnesota, MN (2011) (34)	Images	Photographic facial images for use by maxillofacial surgeons	Composite images created using Adobe Photoshop by blending at least 2 original facial images selected by subjective analysis of facial characteristics and an expected best fit	None of the familiar test faces were initially recognized by the subjects (dental students) after viewing the composite images; subjects rated all composites as clinically realistic looking images; subjects rated composites as a more effective method for de-identification compared to placing black rectangles over eyes and eyebrows	Subjects correctly identified the face 62% of the time when they were primed to viewing a composite containing a named familiar face; methods to make a composite image is non-automated (single facial composite takes 10-15 minutes to create)
Gulcher, deCODE Genetics, Inc., Reykjavik, Iceland (2000) (42)	Biosamples	DNA samples	A population-based list of patients with a particular disease is reversibly encrypted by the Data Protection Commission (DPC) of Iceland that converts SS# to an alphabet-derived character string(PN); the lab runs the PN list against a population-based encrypted genealogy database; a list of patients of interest for genetic studies is sent to the DPC, whom decodes the list and generates a patient list with SS# for the physician; blood samples obtained from interested patients are coded with a temporary number (SN) that is linked to the patient's SS#; a DPC officer encrypts the list of SS-SN to PN-SN and it is sent with the blood to the lab, where the tubes are scanned into a sample storage program and the SN is replaced by relabeling with a permanent in-house sample number (iSN), which remains with the DNA samples isolated from the blood	PN is not used outside the laboratory and direct personal identifiers, such as names or SS#, never enter the laboratory; the only connection between samples or data in the laboratory and the patient or data on the clinical side is through the PN and the DPC, the sole keeper of the encryption code; there remains a method for de-identification if future genetic analyses result in clinically relevant data; the system has been successfully used for 3 years prior to publication, to work with patient lists covering almost 30 common diseases averaging about 2,500 patients per disease along with blood samples from over 35,000 Icelanders	The authors estimate that it takes about 3 hours for a list of 1000 individuals to be encrypted or decrypted, including checking and verifying the personal identifiers, although the authors discuss that a core encryption facility within an institution may increase efficiency
Hara, University of Tokyo, Japan (2003) (43)	Biosamples	DNA samples	Temporary codes are created by an administrator for the sample, informed consent document, and clinical data document; code is printed on a label with a seal covering the number; a researcher	Only the administrator, who is not allowed to be involved in any genetic research conducted at the institution, can link the specimen to the identifying information; anonymizing	The biosamples are not truly anonymized (i.e., what the authors refer to as "unlinked anonymity", or removing all the personal identifying information from a

Primary Author, Location (Year) (Citation)	Type	Test Corpus, Other	Description	Strengths	Limitations
			submits samples and documents to the administrator, who creates a data file in which the identifying and non-identifying data are linked to the codes, and then assigns a permanent, anonymous, random 6-digit number to the codes; an anonymous data file that links this number only to the clinical parameters is created; sample labels are replaced by the anonymous number; researcher receives the anonymous data file and coded specimens; DNA is extracted from anonymous specimens and the researcher has access to the anonymous data file	cost is approximately 40 cents per specimen	specimen without retaining any key and thus preventing any way the specimen can be traced back to the person from whom it was obtained), since the researcher can request the IRB to permit linking of the specimen to personal identifying information
Roden, Vanderbilt University, Nashville, TN (2008) (17)	Biosamples	DNA samples and EHR clinical notes	MRNs on biosamples and linked EHRs are replaced with a 128-character research unique identifiers (RUIs) generated by a one-way SHA, and DNA from the samples are extracted and stored; EHRs linked to biosamples are de-identified by De-Id (see Gupta, 2001)(11), supplemented with preprocessing and postprocessing to scrub PHI; names are scrubbed by complementary use of census-derived name dictionaries and a matching function based on the name information contained in the header files of the original EHRs, as well as use of an institutional employer dictionary; all dates are shifted 1-364 days into the past, and PHI are replaced with corresponding tags	No link is maintained between MRNs and RUIs; the de-identification algorithm removed 5,378 of the 5,472 identifiers, with an error rate for complete HIPAA identifiers of <0.1%; the aggregate error rate (including any potential error, i.e., non-HIPAA items, partial items, and items not inherently related to identity) was 1.7% (95% CI 1.4-2.1%); de-identified records of the first 16,102 samples that were accepted contained a mean of 14 ± 18 ICD-9 codes; in the initial analysis, the error rate for all names contained in the record was 3% (95% CI 2.1-3.8%);	Entire biobank system required significant institutional investment and resources; inability to recontact patients for acquiring additional biosamples
Furuta, National Cancer Center Hospital, Tokyo, Japan (2010) (42)	Biosamples	Bio-repository of post-immunological and various clinical test samples	Blood samples transferred from a clinical ID tagged tube to a repository ID tagged tube (tube transfer procedure for de-identification)	Offers de-identification solution to post-clinical test sample-based biobank storage of a high-volume center (80% of 90,000 samples per year need to be de-identified with a tube transfer procedures); more than 250,000 samples are stored	Risk of mis-transfer of samples due to a manual process; previously collected and frozen samples need to be thawed prior to de-identification procedure, which may affect sample quality

[*] Table is sorted by Type, Year, and Primary Author

Abbreviations and Definitions: CCUW = Clinical and Common Usage Words; CRF = Conditional Random Fields; CT = computed tomography; DICOM = Digital Imaging and Communications in Medicine; DOB = date of birth; EHR = electronic health record; fallout = false positive rate; GATE = General Architecture for Text Engineering; GNU GPL = GNU Project General Public License; HIPAA = Health Insurance Portability and Accountability Act; HL7 = Health Level Seven; MESH = Medical Subject Headings; MR = magnetic resonance; MRN = medical record number; NER = Named Entity Recognition; NLP = Natural Language Processing; OCR = optical character recognition; PHI = protected health information; precision = positive predictivity; RTN = recursive transition networks; recall = sensitivity; SHA = secure hash algorithm; SNOMED = Systematized Nomenclature of Medicine); SPIN = Shared Pathology Informatics Network; SS# = U.S. social security number; UMLS = Unified Medical Language System; XML = extensible markup language