

De-identification of free text data containing personal health information: a scoping review of reviews

Bekelu Negash^{1,*}, Alan Katz^{1,2}, Christine J. Neilson³, Moniruzzaman Moni⁴, Marcello Nesca¹, Alexander Singer², and Jennifer E. Enns¹

Submission History

Submitted:	12/04/2023
Accepted:	22/08/2023
Published:	12/12/2023

¹Manitoba Centre for Health Policy, Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba

²Department of Family Medicine, Rady Faculty of Health Sciences, University of Manitoba

³Neil John Maclean Health Sciences Library, University of Manitoba

⁴George & Fay Yee Centre for Healthcare Innovation, Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba

Abstract

Introduction

Using data in research often requires that the data first be de-identified, particularly in the case of health data, which often include Personal Identifiable Information (PII) and/or Personal Health Identifying Information (PHII). There are established procedures for de-identifying structured data, but de-identifying clinical notes, electronic health records, and other records that include free text data is more complex. Several different ways to achieve this are documented in the literature. This scoping review identifies categories of de-identification methods that can be used for free text data.

Methods

We adopted an established scoping review methodology to examine review articles published up to May 9, 2022, in Ovid MEDLINE; Ovid Embase; Scopus; the ACM Digital Library; IEEE Explore; and Compendex. Our research question was: What methods are used to de-identify free text data? Two independent reviewers conducted title and abstract screening and full-text article screening using the online review management tool Covidence.

Results

The initial literature search retrieved 3,312 articles, most of which focused primarily on structured data. Eighteen publications describing methods of de-identification of free text data met the inclusion criteria for our review. The majority of the included articles focused on removing categories of personal health information identified by the Health Insurance Portability and Accountability Act (HIPAA). The de-identification methods they described combined rule-based methods or machine learning with other strategies such as deep learning.

Conclusion

Our review identifies and categorises de-identification methods for free text data as rule-based methods, machine learning, deep learning and a combination of these and other approaches. Most of the articles we found in our search refer to de-identification methods that target some or all categories of PHII. Our review also highlights how de-identification systems for free text data have evolved over time and points to hybrid approaches as the most promising approach for the future.

Keywords

de-identification; Health Insurance Portability and Accountability Act; electronic medical records; machine learning; personal health information

*Corresponding Author:

Email Address: Bekelu.negash@umanitoba.ca (Bekelu Negash)

Introduction

The production, collection and use of population data for research is becoming more prevalent across multiple sectors, but particularly in health and healthcare [1–3]. For example, the use of electronic health records has seen a significant increase among researchers and clinicians [4, 5]. However, population datasets often contain Personal Identifiable Information (PII) and/or Personal Health Identifying Information (PHII), which researchers have the responsibility to keep confidential. In Canada, the use of population data containing PII and PHII in research is governed by the Canadian Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans, which includes three core principles: respect for persons, concern for welfare, and justice [6]. One effective way to preserve privacy and abide by this ethical framework is to de-identify the data before it is used in research. *De-identification* refers to the removal or masking of PII/PHII in a dataset and for research purposes, it may be preferable to anonymisation, a process that eliminates all identifying details in a data record with no way of back-tracking to link related data records together [7]. When a record number, file number or other encrypted linkage tool is retained in the original data, the data are not referred to as ‘anonymised’ but are instead ‘de-identified’ and can be used in data linkage applications.

The federally mandated Freedom of Information and Protection of Privacy Act (FIPPA) and the provincially mandated Personal Health Information Act (PHIA) provide definitions of PII and PHII and set out guidelines to inform the process of de-identifying structured data [7]. *Structured data* are organised into specific value sets and are typically stored in a database [8]. Meanwhile, *unstructured* or *free text data* do not have pre-defined values; for example, reports created by physicians may contain free text data that vary widely in structure and content [8]. Currently, there is very little formal guidance available on how to de-identify free text data, and none that we could find that differentiates PII and PHII in the de-identification process. In this matter, the distinction between PII (recorded information that could identify an individual or groups of individuals) and PHII (specific health information about an individual or groups of individuals) is important, because specific approaches for de-identification are needed if health information is present in the data [8] (see Table 1 for examples [9]).

The natural language processing (NLP) research community has made great strides in developing methods for automatically de-identifying data. There are currently two primary approaches in use:

1. **Rule-based methods** use pattern-matching with set conditions to satisfy a rule [7]. A rule-based method could, for example, be used to find names, addresses, or email addresses in data records. Advantages to rule-based methods include that they are relatively simple to create and do not require labelled data [10], and certain sub-types (e.g., generalisation, suppression, and data perturbation) can also be used to prevent individual records from being traced back and re-identified, if this is an important aspect of the research study [10]. However, developing rules can be time-consuming since it is

difficult to include all possible examples in the rules, and the experts who design the rules may make assumptions about the data that could limit the effectiveness of the de-identification process [11].

2. **Machine learning (ML)/statistical learning methods** use probabilistic or classification modelling to describe the structure of data or generate predictions based on inputs from a dataset. Machine/statistical learning algorithms are classified as either supervised or unsupervised. Supervised ML requires that a sample of data be labelled manually to support the model. The advantage of supervised ML approaches is that they automatically learn sophisticated pattern recognition [12]. However, it can be more difficult to identify sources of error in unsupervised deep learning ML model than in a rule-based approach. In addition, when it comes to rare types of information, ML methods can also have a lower performance compared to rule-based methods [11].

These two methods can be used for de-identification of both structured and free text data. De-identifying structured data is a relatively straight-forward process compared to de-identifying free text data, because structured data typically have a limited number of clearly identified fields; in addition, there is some literature to inform and guide the process [7]. De-identifying free text data, however, may necessitate a more sophisticated approach, since identifying information may occur anywhere in the free text and may include either PII or PHII or both. Some researchers have developed hybrid approaches in an attempt to combine the advantages of rule-based and ML methods for de-identification of PHII [11]. The hybrid approaches take advantage of the fact that certain types of PHII exhibit predictable lexical patterns and thus lend themselves well to de-identification via rule-based methods, whereas other frequently encountered PHII types, particularly those with unpredictable lexical variations, are more amenable to machine learning approaches [13]. More recently, the use of deep learning methods has been explored to de-identify electronic health records [14]. These methods have the ability to learn the most relevant features from the raw data, minimising the need for human input and making the pre-processing and feature engineering steps less time consuming [15].

Data de-identification techniques are advancing quickly and have a growing number of applications in research settings. In this scoping review, we provide an overview of what is known about NLP methods used to de-identify free text data.

Methods

We based our scoping review approach on Arksey and O'Malley's (2005) well-established scoping review framework, which comprises five stages [16]: identifying the research question; identifying relevant studies; study selection; charting the data; and collating, summarising, and reporting the results. Our research question was: What methods are used to de-identify free text data?

Table 1: Examples of personal identifiable information and personal health identifying information

Personal Identifiable Information (PII):

- Name, contact information
- Age, sex, sexual orientation, marital or family status
- Ancestry, race, colour, nationality, national or ethnic origin
- Religion, creed, religious belief, association, or activity

Personal Health Identifying Information (PHII):

- An individual's health or health care history, including genetic information about the individual
- The provision of health care to the individual
- Payment for health care provided to the individual, including personal health information number (PHIN) and any other identifying number, symbol or particular assigned to an individual
- Any identifying information about the individual collected in the course of, and incidental to, the provision of health care or payment for health care

Search strategy

A professional librarian and research coordinator developed the search strategy. Initially, one search strategy was tailored to the health database Ovid MEDLINE, and another was tailored to the ACM Digital Library, a computing literature database. Both strategies were independently peer-reviewed according to the Peer Review of Electronic Search Strategies (PRESS) checklist by a second librarian with the required subject specialisation [17]. The final search strategies were translated for use in Ovid Embase; Scopus; IEEE Explore; and Compendex. All searches were conducted on May 9, 2022. No date limits were used. The MEDLINE and Embase searches were limited to English language publications. Complete search histories for each database are available online (<http://hdl.handle.net/1993/37168>).

Article screening and selection

Using the study selection criteria in Table 2, two independent reviewers examined the titles and abstracts of the search results. Articles that were ambiguous were discussed with the research coordinator and a consensus decision was reached on whether or not to include them in full-text article screening. The two reviewers then completed full-text article screening on the selected articles.

Data extraction and analysis

The data categories we extracted are presented in Table 3. We analysed and summarised the results in accordance with

the PRISMA-ScR reporting checklist [18]. The data analysis was designed to provide an overview of methods used for de-identifying free text data.

Results**Article screening and selection**

As shown in Figure 1, we identified 3,312 articles in the initial search, and removed 329 duplicates. The two reviewers had a 95.5% agreement rate during title and abstract screening; 4.2% (124) of the initial search results were included at this stage. After full-text article screening, 14.5% (18) of the 124 articles met the specified criteria and were included in the scoping review.

Article characteristics

Of the 18 included articles, twelve were from the computer science literature. Most (83%) were literature reviews. The other information we planned to extract was scantily available – only three of the 18 articles mentioned the databases and registers the authors used for their searches, two articles provided information regarding the year of publication of the primary articles, number of articles or the percentage of the articles included in their review, and another three articles indicated what inclusion/exclusion criteria the authors used. Table 4 presents more details on these latter articles.

Table 2: Study selection criteria

Inclusion criteria	Exclusion criteria
• Studies published in English • Discusses methods of de-identification • Focused on free text data • Review article	• Focused on the accuracy and representability of the text after de-identification • Focused on privacy and less on the method of de-identification • Used de-identified text for data • Focused on cryptography de-identification methods

Table 3: Data categories extracted

Article Information

- Journal discipline (medicine, computer science, both, other)
- Type of review
- Publication year range of articles included in the review
- Inclusion/exclusion criteria that the review article used
- Number of articles cited in the review article

Any mention of legal framework or guidelines

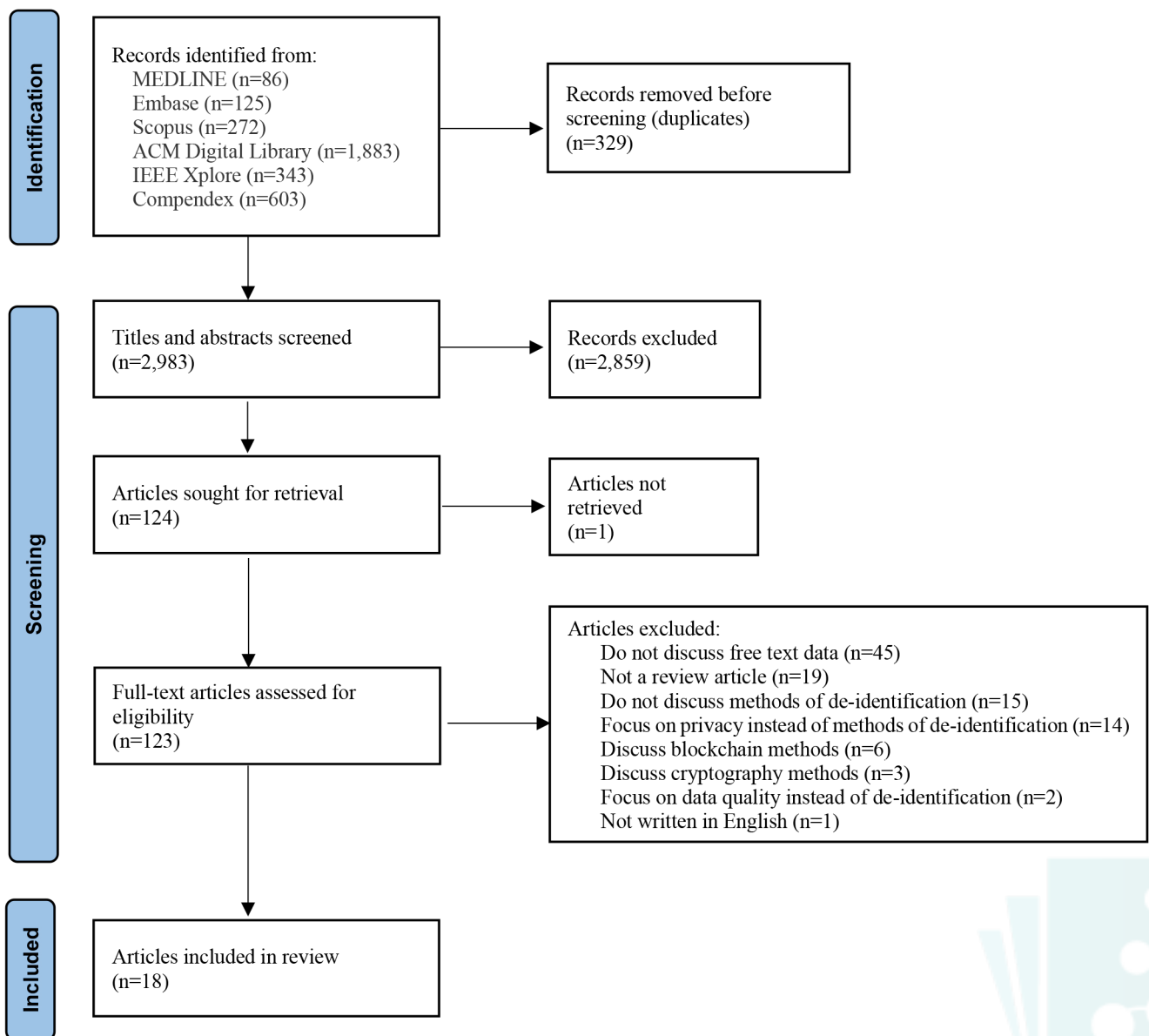
Type of PII/PHI addressed

Type of text data (medical [e.g., EHR, safety reports], social media, other)

De-identification methods

Evaluation metrics for the de-identification outcome

Figure 1: PRISMA diagram – article search and selection process



Legal framework or guidelines mentioned

Of the 18 included articles, 12 mentioned legal acts governing data privacy. The Health Insurance Portability and

Accountability Act (HIPAA) was mentioned in nine articles (75%), and four (33%) of the articles considered the European Union's General Data Protection Regulation (GDPR). The remaining legal acts mentioned were from Canada, China,

Table 4: Characteristics of the articles that reported inclusion and exclusion criteria

Author	Databases/ registries searched	Type of review	Year range	Inclusion criteria	Exclusion criteria	Number of articles included
Meystre et al., 2010 [19]	PubMed, conference proceedings, ACM Digital Library	Literature review	1995–2010	Key terms: de-identification, anonymisation, text scrubbing, narrative text, and/or automated text de-identification. For the ACM Digital Library, the same terms were used, with the addition of medical, medicine, biomedical or clinical.	Focused on structured data, radiological or face image de-identification, manual de-identification	18
Shickel et al., 2017 [15]	Google Scholar	Literature review	Up to August 2017	Key terms: Electronic health records (EHRs) or electronic medical records (EMR) in conjunction with deep learning or a specific deep learning method (e.g., recurrent neural network [RNN]).	Not described.	44 articles on privacy-preserving methods, including cryptography-based methods, and approximately 4 articles on anonymisation methods of de-identification (exact number not specified in article).
Kushida et al., 2012 [20]	BIOSIS Previews, CINAHL, Inspec, MEDLINE, SciVerse, Scopus, Web of Science	Systematic review	Up to June 30, 2011	Key terms: De-identify, de-identification, anonymise, anonymisation, data scrubbing, and text scrubbing. Reviewed additional articles extracted from references of articles from search.	Citations for non-relevant article types (e.g., reviews, opinions, editorials, or commentaries), outside medical records domain, de-identification, or anonymisation strategy lacked sufficient detail to understand or interpret it.	45

Australia and New Zealand – see Table 5 for more details.

Types of PII and PHII

Seventeen articles mentioned different types of PII and PHII (Table 6). Eight of these articles identified methods that de-identified protected health information according to all 18 categories of HIPAA [15, 19–21, 23, 24, 26, 29] while others identified some of the HIPAA categories (Table 7).

Types of free text data

Of the 18 articles, eight examined free text health data from electronic health records [15, 19–21, 23–26]. Another

eight articles mentioned big data [12, 28, 30–35] but did not elaborate further on data type. Two articles mentioned both health data and big data [22, 27].

Methods of de-identification for free text data

The de-identification approaches for free text data we found in the literature can be categorised into four overlapping groups: rule-based methods, ML methods, deep learning (a subset of machine learning) methods, and hybrid methods. The non-automated rule-based learning approaches used are summarised in Table 8, and all other de-identification approaches and system/software packages mentioned are presented in Table 9.

Table 5: Legal frameworks or guidelines referred to in the articles

Legal framework or guidelines mentioned	Article(s)
Health Insurance Portability and Accountability Act (HIPAA)	[12, 15, 19–25, 25–28]
General Data Protection Regulation (GDPR)	[12, 22, 23, 26]
Health Information Technology for Economic and Clinical Health (HITECH) Act	[21, 23]
Personal Information Protection and Electronic Documents Act (PIPEDA)	[12, 23]
Consumer Data Right	[23]
China Civil Code	[23]
Medical Practitioners Act	
Personal Information Protection Law	[23]
Regulations on Medical Records Management In Medical Institutions	[23]
Children's Online Privacy Protection Act	[12]
Genetic Information Non-discrimination Act	[21]
Gramm-Leach-Bliley Act	[12]
Health Information Privacy Code	[26]

Table 6: Types of PII/PHI referred to in the articles

Types of PII/PHI	Article(s)
Individuals' identifiers (such as credit card records) and interaction privacy (e.g., use of voice/fingerprint)	[30]
Key attributes (e.g., ID, name, social security), quasi-identifiers (e.g., birth date, zip code, position, job, blood type), sensitive attributes (e.g., salary, medical examinations, credit card releases)	[12, 31–33]
7 types of PHII, including personal names, ages, geographical locations, hospitals and healthcare organisations, dates, contact information, IDs	[19]
PHI: patient name, phone number, physician name, medical history. PII: names, addresses, contact numbers	[22]
18 categories of PHI according to HIPAA, quasi-identifiers, 9 categories of personal information according to China Civil Code (name, birthday, ID number, biometric information, home address, phone number, email address, health condition information, and personal tracking information)	[23]
PHI according to HIPAA, doctor's name and years extracted from dates	[20]
Direct identifiers (e.g., name, mailing address, email, social security number, phone number or driver's license number) and indirect identifiers (e.g., birth date, postal code, and sex)	[27, 28]

Table 7: HIPAA categories

- Names
- All geographical subdivisions smaller than a state except the first two digits of the zip code
- All elements of dates (except year)
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers or serial numbers, including plate numbers
- Device identifiers or serial numbers
- Web URLs
- Internet protocol addresses
- Biometric identifiers
- Full-face photographs and comparable images
- Any other unique identifying number, characteristic, or code

Nine articles referred to methods like rule-based automated learning, i.e., methods created to de-identify text data automatically using HMS Scrubber, an open-source de-identification tool that employs a three-step process to remove PHII from medical documents [36], and DE-ID rule-based automated system that uses sets of rules, pattern-matching algorithms, and dictionaries to identify PHII in medical documents [19–21]. Machine learning approaches such as MIST (MITRE Identification Scrubber Toolkit, software that uses samples of de-identified text that enable it to learn contextual features that are necessary for accuracy) were mentioned in four articles, [19–21, 24] the Health Information De-identification (HIDE) system was mentioned in two articles [19, 20].

System/software packages containing de-identification methods can also be further divided into specific heuristic,

pattern-based and statistical learning-based systems. The systems based on deep learning use a combination of specific de-identification approaches. Some articles also mentioned hybrid systems that achieved outstanding results in various natural language processing challenges pertaining to de-identification. For example, systems developed for the 2014 i2b2 challenge is a hybrid system based on machine learning and rule-based methods [13, 37–39].

Evaluation metrics

The metrics mentioned to measure performance by the articles are presented in Table 10. Six out of the 18 articles mentioned evaluation metrics for assessing the performance of NLP de-identification approaches. Some articles used terms

Table 8: Rule-based de-identification approaches in the included articles

De-Identification process	Rule-based learning methods	Article(s)
Anonymisation Models	• K-anonymity, l-diversity, t-closeness and M-variance • β-likeness and suppression • Cluster-based Missing and Value Imputation • Differential privacy • Fuzzy-based (clustering)	[12, 23, 30–35] [12] [34] [28] [27]
Data Perturbation	• Value-based (e.g., uniform perturbation, probability distribution/randomisation) • Dimension-based (e.g., random rotation transformation, random projection) • Randomisation	[27, 30, 32] [27, 31] [23, 27, 33]

Table 9: Additional categories of de-identification approaches in the included articles, including systems and software

Rule-Based automated	Machine learning	Deep learning	Hybrid
• DE-ID [40–42] • National Library of Medicine Scrubber [43, 44] • Privacy Analytics Risk Assessment Tool [45] • HMS Scrubber [36, 46–48] • MEDTAG [49, 50] • Regenstrief Institute System [51] • Concept-Match [52, 53] • VA system [54] • Encryption Broker Software [56] • Medical information anonymisation [57] • N-Sanitisation [58] • Medical De-identification System (MEDs) [59, 60] • MedLEE [42, 61] • Deid-Swe [62]	• Software based on machine learning for the CEGS N-GRID 2016 de-id shared task [55] • MIST (Identification Scrubber Toolkit) [63, 64] • Stat De-id [65] • UCLA system [66] • System for the 2006 i2b2 de-identification challenge (based on the Conditional Random Field [67] • HIDE [68] • System for the 2006 i2b2 de-identification challenge (based on Support Vector Machine method) [25, 69–71] • System for the 2006 i2b2 de-identification challenge (based on Decision Tree method) [72, 73] • Health Information DE-identification [74] • Hidden Markov Models based tagger [71, 75] • HitzalMed [42] • Hidden Markov Model using Dirichlet Process [71] • Systems based on MALLET and conditional random field (CRF) [76]	• NeuroNER [64] • System based on Bi-directional Long Short-Term Memory [47, 64] • Frequency-filtering-based system [53, 54] • Systems based on Bidirectional Encoder Representations from Transformers and Multilingual Bidirectional Encoder Representations from Transformers [59] • Systems based on two variants (Elman and Jordan) of RNN [74] • Text Skeleton-Recurrent Neural Network (Combination of RNN and text skeleton) [74] • Transfer learning with RNN [77]	• System for the 2014 i2b2 de-identification challenge [37, 54, 73] • System based on CRF and Bi-LSTM [48, 49, 64] • System based on the combination of convolutional neural network, Bi-LSTM, and CRF [78] • System for the 2014 i2b2 de-identification challenge (based on combination of CRF and rule-based approaches) [13, 37–39] • System for the 2016 i2b2 de-identification challenge [13, 20, 58] • Multilevel Hybrid Semi-Supervised Learning Approach (MLHSLA) [62] • System based on mDEID and CiDEID [79] • System for the 2016 i2b2 de-identification challenge (based on Bi-LSTM, CRF, and rule-based approaches) [80] • System based on Bi-LSTM and human-engineered features from EHRs [81]

commonly used in the computer science literature such as *recall* and *precision* while others used terms that have the same meaning from epidemiology such as *sensitivity* and *specificity*. Additionally, while the articles discuss the same metrics, some of them use different formulas in varying contexts. For instance, in Kushida et al. (2012), the term *precision* is employed to evaluate the performance of Stat De-id, a statistical learning-based system originally introduced in Uzuner et al. (2008) [20, 65]. However, in Meystre et al.

(2010), the *precision* formula is not provided, but instead, reference is made to how HMS Scrubber was evaluated by Beckwith et al. (2006) [19, 36].

Discussion

Free text data contain a wealth of information that is valuable in research. To take full advantage of this information,

Table 10: NLP metrics mentioned in the included articles

Evaluation metric	Articles
• Precision	[19, 20, 24, 25]
• Accuracy	[15, 20]
• Area under ROC curve	[15, 20]
• Sensitivity	[15, 20]
• F-measure	[15, 19, 20, 24–26]
• Recall	[15, 19, 20, 24, 25]
• Specificity	[19, 20]

de-identification approaches for free text data must ensure the privacy and confidentiality of individuals described in the data. The discussion of de-identification of data in health research previously focused on structured data. The growth and importance of free text data in health records and health research has resulted in the need for advances in de-identification approaches. This scoping review of reviews identifies published de-identification methods for free text data. We have categorized the methods as rule-based methods, machine learning, deep learning and a combination of these and other approaches. Most of the articles we found in our search refer to de-identification methods (primarily rule-based and machine learning methods) that target some or all categories of PHII defined by HIPAA.

In general, experts in the field are using rule-based methods with anonymisation models to de-identify data; in particular, they use K-anonymity, I-diversity and t-closeness. Sakpere et al. (2014) assert that K-anonymity methods are best suited for data stream anonymity, such as phone numbers [31]. However, Senosi et al. (2017) found that researchers only give anonymisation strategies an average rating for protecting privacy [32]. Additionally, Stubbs et al. (2015) observes that even if automated rule-based solutions are beneficial, some PHII is still included in the data since the success of the de-identification process depends on the dictionaries used [25]. Yogarajan et al. (2020) argues that machine learning methods for de-identification need to improve in areas such as maintaining correctness and usability of data [26]. Meystre et al. (2010) states that machine learning methods combined with rule-based approaches such as HMS Scrubber perform better than a single method at de-identification of free text data [19].

Recently published articles reviewed a number of approaches, including systems based on machine learning and hybrid systems that use a combination of different de-identification methods, including deep learning methods (e.g., NeuroNER and Bidirectional Encoder Representations from Transformers (BERT) [22, 26]. Shickel et al. (2018) found systems based on deep learning performed better than other methods on lexical features [15]. However, deep learning techniques require large datasets to perform effectively [15]. Deep learning methods also make validating accuracy challenging due to the nature of the method. While they do represent significant progress in de-identification, the size of the required datasets for acceptable performance is an important limitation.

Conclusion

This scoping review provides an overview of de-identification methods for free text data. As computation power and the availability of free text from electronic health records have increased, the importance of de-identification methods in advancing the use of text data for research has also grown. While this review sought to classify de-identification techniques, no single approach or rule-based method was found to meet the high standards required to address the needs of research privacy regulators in protecting the privacy of patients since no single approach could reliably de-identify all PHII in population data records [20]. The combination of multiple tools in a hybrid format appears to be the most promising future direction.

Ethics

The University of Manitoba Health Research Ethics Board does not require review of review articles.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, and/or publication of this article.

Acknowledgements

We thank Li Zhang, MLIS (University of Saskatchewan Library) for peer review of the *ACM Digital Library* search strategy. The librarian who peer reviewed the MEDLINE search strategy does not wish to receive formal acknowledgement, but her contribution to this review is equally valued.

Funding

This research was supported by a foundation grant from the Canadian Institutes of Health Research (Foundation grant reference number 148427).

References

1. Ngiam KY, Khor IW. Big Data and Machine Learning Algorithms for Health-Care Delivery. *Lancet Oncol* [Internet] 2019 May [cited 2023 25];20:e262–73. [https://doi.org/10.1016/S1470-2045\(19\)30149-4](https://doi.org/10.1016/S1470-2045(19)30149-4)
2. Tao D, Yang P, Feng H. Utilization of Text Mining as a big Data Analysis Tool for Food Science and Nutrition. *Compr Rev Food Sci Food Saf* [Internet] 2020 Mar. [cited 2023 25];19:875–94. <https://doi.org/10.1111/1541-4337.12540>
3. Rudrapatna VA, Butte AJ. Opportunities and Challenges in Using Real-World Data for Health Care. *J Clin Invest* [Internet] 2020 Feb. [cited 2023 25];130:565–74. <https://doi.org/10.1172/JCI129197>

4. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The Evolving Use of Electronic Health Records (EHR) for Research. *Semin Radiat Oncol* 2019 Oct.;29:354–61. <https://doi.org/10.1016/j.semradonc.2019.05.010>
5. Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, et al. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Comput Surv [Internet]* 2022 [cited 2023 25];55:33. Available from: <https://doi.org/10.1145/3490234>.
6. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans [Internet]. Government of Manitoba 2018 [cited 2023 2]; Available from: www.nserc-crsng.gc.ca.
7. Office of the Information and Privacy Commissioner of Ontario. De-identification Guidelines for Structured Data [Internet]. Information and Privacy Commissioner of Ontario 2016 [cited 2023 28]; Available from: <https://www.ipc.on.ca/resource/de-identification-guidelines-for-structured-data/>.
8. Abu-El-Rub N, Urbain J, Kowalski G, Osinski K, Spaniol R, Liu M, et al. Natural Language Processing for Enterprise-scale De-identification of Protected Health Information in Clinical Notes. [cited 2023] Available from: <https://pubmed.ncbi.nlm.nih.gov/35854742/>.
9. University of Manitoba. Access and Privacy [Internet]. Access and Privacy Office 2023 [cited 2023 28]; Available from: <https://umanitoba.ca/access-and-privacy/>.
10. Golemund G, Wickham H. R for data science [Internet]. 2019. Available from: <https://r4ds.had.co.nz/>.
11. Lee HJ, Wu Y, Zhang Y, Xu J, Xu H, Roberts K. A Hybrid Approach to Automatic De-identification of Psychiatric Notes. *J Biomed Inform [Internet]* 2017 Nov. [cited 2023 2];75S:S19–27. <https://doi.org/10.1016/j.jbi.2017.06.006>
12. Basso T, Matsunaga R, Moraes R, Antunes N. Challenges on Anonymity, Privacy, and big Data. In: *Proceedings - 7th Latin-American Symposium on Dependable Computing, LADC 2016*. 2016. <https://doi.org/10.1109/LADC.2016.34>
13. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic De-identification of Electronic Medical Records Using Token-level and Character-Level Conditional Random Fields. *J Biomed Inform* 2015 Dec.;58:S47–52. <https://doi.org/10.1016/j.jbi.2015.06.009>
14. Yang X, Lyu T, Li Q, Lee CY, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med Inform Decis Mak [Internet]* 2019 Dec. [cited 2023 30];19:1–9. <https://doi.org/10.1109/ICHI.2019.8904544>
15. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *Institute of Electrical and Electronics Engineers Journal of Biomedical and Health Informatics* 2018;22. <https://doi.org/10.1109/JBHI.2017.2767063>
16. Arksey H, O'Malley L. Scoping Studies: Towards a Methodological Framework. *International Journal of Social Research Methodology: Theory and Practice* 2005;8. <https://doi.org/10.1080/1364557032000119616>
17. McGowan J, Sampson M, Salzweid DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol* 2016 Jul.;75:40–6. <https://doi.org/10.1016/J.JCLINEPI.2016.01.021>
18. Tricco AC, Lillie E, Zarin w, O'Brian KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018;169. <https://doi.org/10.7326/M18-0850>
19. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic De-identification of Textual Documents in the Electronic Health Record: A Review of Recent Research. *BMC Med Res Methodol* 2010;10. <https://doi.org/10.1186/1471-2288-10-70>
20. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for De-identification and Anonymization of Electronic Health Record Data for use in Multicenter Research Studies. *Med Care* 2012;50. <https://doi.org/10.1097/MLR.0b013e3182585355>
21. Kayaalp M. Patient Privacy in the era of big Data. *Balkan Med J* 2018;35. <https://doi.org/10.4274/balkanmedj.2017.0966>
22. Mahendran D, Luo C, McInnes BT. Review: Privacy-Preservation in the Context of Natural Language Processing. *Institute of Electrical and Electronics Engineers Access [Internet]* 2021 [cited 2023 10];9. Available from: <https://doi.org/10.1109/ACCESS.2021.3124163>.
23. Xiang D, Cai W. Privacy Protection and Secondary Use of Health Data: Strategies and Methods. *Biomed Res Int* 2021;2021. <https://doi.org/10.1155/2021/6967166>
24. Stubbs A, Filannino M, Uzuner Ö. De-identification of Psychiatric Intake Records: Overview of 2016 CEGS N-GRID shared tasks Track 1. *J Biomed Inform* 2017;75. <https://doi.org/10.1016/j.jbi.2017.06.011>
25. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *J Biomed Inform* 2015;58. <https://doi.org/10.1016/j.jbi.2015.06.007>
26. Yogarajan V, Pfahringer B, Mayo M. A Review of Automatic end-to-end De-Identification: Is High Accuracy

- the Only Metric? Applied Artificial Intelligence 2020;34.
<https://doi.org/10.1080/08839514.2020.1718343>
27. Zainab SS, Kechadi T. Sensitive and private data analysis: A systematic review. In: ACM International Conference Proceeding Series. 2019.
<https://doi.org/10.1145/3341325.3342002>
28. Youm HY. An Overview of De-identification Techniques and Their Standardization Directions. Institute of Electronics, Information and Communication Engineers Transactions on Information and Systems 2020;E103D. <https://doi.org/10.1587/transinf.2019IC10002>
29. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015;58.
<https://doi.org/10.1016/j.jbi.2015.06.007>
30. Binjubeir M, Ahmed AA, Ismail MA Bin, Sadiq AS, Khurram Khan M. Comprehensive Survey on big Data Privacy Protection. Institute of Electrical and Electronics Engineers. Access 2020;8:20067–79.
<https://doi.org/10.1109/ACCESS.2019.2962368>
31. Sakpere AB, Kayem AVDM. A State-of-the-art Review of Data Stream Anonymization Schemes. In: Information Security in Diverse Computing Environments. 2014.
<https://doi.org/10.4018/978-1-4666-6158-5.ch003>
32. Senosi A, Sibiya G. Classification and Evaluation of Privacy Preserving Data Mining: A review. In: Institute of Electrical and Electronics Engineers AFRICON: Science, Technology and Innovation for Africa. 2017.
<https://doi.org/10.1109/AFRCON.2017.8095593>
33. Shanthi AS, Karthikeyan M. A Review on Privacy Preserving Data Mining. 2012 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2012 2012;
<https://doi.org/10.1109/ICCIC.2012.6510302>
34. Deng H, Wang Z, Zhang Y. Overview of Privacy Protection Data Release Anonymity Technology. International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) Conference on Intelligent Data and Security (IDS) 2021 May;151–6.
<https://doi.org/10.1109/BigDataSecurityHPSCIDS52275.2021.00037>
35. Shelake VM, Shekokar N. A Survey of Privacy Preserving Data Integration. In: International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICECCOT . 2017.
<https://doi.org/10.1109/ICECCOT.2017.8284559>
36. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and Evaluation of an Open Source Software Tool for De-identification of Pathology Reports. BMC Med Inform Decis Mak 2006;6.
<https://doi.org/10.1186/1472-6947-6-12>
37. He B, Guan Y, Cheng J, Cen K, Hua W. CRFs Based De-identification of Medical Records. J Biomed Inform 2015 Dec.;58:S39–46.
<https://doi.org/10.1016/J.JBI.2015.08.012>
38. Yang H, Garibaldi JM. Automatic Detection of Protected Health Information From Clinic Narratives. J Biomed Inform [Internet] 2015 Dec. [cited 2023 2];58 Suppl:S30–8.
<https://doi.org/10.1016/j.jbi.2015.06.015>
39. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining Knowledge- and Data-Driven Methods for De-identification of Clinical Narratives. J Biomed Inform [Internet] 2015 Dec. [cited 2023 2];58 Suppl:S53–9. <https://doi.org/10.1016/j.jbi.2015.06.029>
40. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated De-identification of Free-Text Medical Records. BMC Med Inform Decis Mak 2008;17. <https://doi.org/10.1186/1472-6947-8-32>
41. 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.
<https://doi.org/10.1309/e6k3-3gbp-e5c2-7fyu>
42. Lima S, Perez N, García-Sardiña L, Sardiña S, Cuadros M. HitzalMed: Anonymisation of Clinical Text in Spanish [Internet]. In: Twelfth Language Resources and Evaluation Conference. 2020 [cited 2023 2]. p. 7038–43. Available from: <https://aclanthology.org/2020.lrec-1.870>.
43. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium [Internet] 2014 [cited 2023 10];2014. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419982/>.
44. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The Pattern of Name Tokens in Narrative Clinical Text and a Comparison of Five Systems for Redacting Them. Journal of the American Medical Informatics Association 2014;21.
<https://doi.org/10.1136/amiajnl-2013-001689>
45. Privacy Analytics Data Anonymization Solution. PARAT Maintenance and Support Information |Privacy Analytic's Privacy and Confidentiality KnowledgeBase [Internet]. 2023 [cited 2023 2]; Available from: <http://knowledgebase.privacy-analytics.com/index.php?/article/AA-00335/0/PARAT-Maintenance-and-Support-Information.html>.
46. Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. Proceedings: a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA

- Fall Symposium [Internet] 1996 [cited 2023 10]; Available from: <https://pubmed.ncbi.nlm.nih.gov/8947683/>.
47. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of Patient Notes With Recurrent Neural Networks. *Journal of the American Medical Informatics Association* [Internet] 2017 May[cited 2023 2];24:596–606. <https://doi.org/10.1093/jamia/ocw156>
 48. Catelli R, Casola V, De Pietro G, Fujita H, Esposito M. Combining Contextualized Word Representation and Sub-document Level Analysis Through Bi-LSTM+CRF Architecture for Clinical De-identification. *Knowledge-Based System* 2021;213. <https://doi.org/10.1016/j.knosys.2020.106649>
 49. Jiang Z, Zhao C, He B, Guan Y, Jiang J. De-identification of Medical Records Using Conditional Random Fields and Long Short-term Memory Networks. *J Biomed Inform* [Internet] 2017 Nov. [cited 2023 2];75S:S43–53. <https://doi.org/10.1016/j.jbi.2017.10.003>
 50. Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical Document Anonymization with a Semantic Lexicon. *PMC- Proceedings AMIA Symposium* [Internet] 2000 [cited 2023 2];729–33. Available from: <https://pubmed.ncbi.nlm.nih.gov/11079980/>.
 51. Thomas SM, Mamlin B, Schadow G, McDonald C. A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. *Proceedings of the AMIA Symposium* [Internet] 2002 [cited 2023 3];777. Available from: <https://pubmed.ncbi.nlm.nih.gov/12463930/>.
 52. Berman JJ. Concept-Match Medical Data Scrubbing: How Pathology Text can be Used in Research. *Arch Pathol Lab Med* 2003;127. <https://doi.org/10.5858/2003-127-680-CMDS>
 53. Li D, Rastegar-Mojarad M, Elayavilli RK, Wang Y, Mehrabi S, Yu Y, et al. A Frequency- Filtering Strategy of Obtaining PHI-Free Sentences From Clinical Data Repository [Internet]. In: *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. Association for Computing Machinery, Inc; 2015 [cited 2023 2]. p. 315–24. <https://doi.org/10.1145/2808719.2808752>
 54. Sadat MN, Aziz MM Al, Mohammed N, Pakhomov S, Liu H, Jiang X. A Privacy-Preserving Distributed Filtering Framework for NLP Artifacts. *BMC Med Inform Decis Mak* [Internet] 2019 Sep. [cited 2023 2];19:1–10. <https://doi.org/10.1186/s12911-019-0867-z>
 55. Bui DDA, Wyatt M, Cimino JJ. The UAB Informatics Institute and 2016 CEGS N-GRID De-identification Shared Task Challenge. *J Biomed Inform* 2017 Nov.;75:S54–61. <https://doi.org/10.1016/j.jbi.2017.05.001>
 56. Pestian JP, Itert L, Andersen C, Duch W. Preparing Clinical Text for Use in Biomedical Research. *Journal of Database Management* [Internet] 2006 Jan. [cited 2023 2]. <https://doi.org/10.4018/jdm.2006040101>
 57. Grouin C, Rosier A, Dameron O, Zweigenbaum P. Testing Tactics to Localize De-Identification. *Stud Health Technol Inform* [Internet] 2009 [cited 2023 3];150:735–9. <https://doi.org/10.3233/978-1-60750-044-5-735>
 58. Iwendi C, Moqurrah SA, Anjum A, Khan S, Mohan S, Srivastava G. N-Sanitization: A Semantic Privacy-Preserving Framework for Unstructured Medical Datasets. *Comput Commun* 2020;161. <https://doi.org/10.1016/j.comcom.2020.07.032>
 59. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. *Appl Soft Comput* 2020 Dec.;97:106779. <https://doi.org/10.1016/j.asoc.2020.106779>
 60. Friedlin FJ, McDonald CJ. A Software Tool for Removing Patient Identifying Information From Clinical Documents. *J Am Med Inform Assoc* [Internet] 2008 Sep. [cited 2023 2];15:601–10. <https://doi.org/10.1197/jamia.M2702>
 61. Morrison FP, Li L, Lai AM, Hripcsak G. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *J Am Med Inform Assoc* [Internet] 2009 Jan. [cited 2023 2];16:37–9. <https://doi.org/10.1197/jamia.M2862>
 62. Velupillai S, Dalianis H, Hassel M, Nilsson GH. 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* [Internet] 2009 Dec. [cited 2023 2];78. <https://doi.org/10.1016/j.ijmedinf.2009.04.005>
 63. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly Retargetable Approaches to De-identification in Medical Records. *Journal of the American Medical Informatics Association* 2007;14. <https://doi.org/10.1197/jamia.M2435>
 64. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an Easy-to-use Program for Named-Entity Recognition Based on Neural Networks [Internet]. In: *Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics (ACL); 2017 [cited 2023 2]. p. 97–102. <https://doi.org/10.18653/v1/D17-2017>
 65. Uzuner Ö, Sibanda TC, Luo Y, Szolovits P. A De-identifier for Medical Discharge Summaries. *Artif Intell Med* [Internet] 2008 Jan. [cited 2023 2];42:13–35. <https://doi.org/10.1016/j.artmed.2007.10.001>
 66. Taira RK, Bui AAT, Kangarloo H. Identification of Patient Name References Within Medical Documents Using Semantic Selectional Restrictions. *Proceedings of the AMIA Symposium* [Internet] 2002 [cited 2023 3];757. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244274/>.

67. Aramaki E, Imai T, Miyo K, Ohe K. Automatic De-identification by Using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data [Internet] 2006; . Available from: <https://www.i2b2.org/NLP/DataSets/Publications.php>.
68. Gardner J, Xiong L. HIDE: An Integrated System for Health Information DE-identification [Internet]. In: International Symposium on Computer-Based Medical Systems. 2008 [cited 2023 2]. p. 254–9. <https://doi.org/10.1109/CBMS.2008.129>
69. Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple M. Identifying Personal Health Information Using Support Vector Machines. [Internet]. In: i2b2 workshop in challenges in natural language processing for clinical data 2006. 2006 [cited 2023 10]. Available from: <https://www.i2b2.org/NLP/DataSets/Publications.php>.
70. Hara K. Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge. [Internet]. In: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. Washington: 2006 [cited 2023 10]. Available from: <https://www.i2b2.org/NLP/DataSets/Publications.php>.
71. Chen T, Cullen RM, Godwin M. Hidden Markov Model Using Dirichlet Process for De-identification. J Biomed Inform 2015 Dec.;58:S60–6. <https://doi.org/10.1016/j.jbi.2015.09.004>
72. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. Journal of the American Medical Informatics Association 2007;14. <https://doi.org/10.1197/j.jamia.M2441>
73. Torii M, Fan J, Yang W, Lee T, Wiley M, Zisook D. De-Identification and Risk Factor Detection in Medical Records [Internet]. In: Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data. Washington: 2014 [cited 2023 10]. https://www.i2b2.org/NLP/HeartDisease/assets/i2b2_2014_schedule_revised.pdf
74. Yadav S, Ekbal A, Saha S, Bhattacharyya P. Deep Learning Architecture for Patient Data De-identification in Clinical Records [Internet]. In: Clinical Natural Language Processing. 2016 [cited 2023 2]. p. 32–41. Available from: <https://aclanthology.org/W16-4206>.
75. Medlock B. An Introduction to NLP-based Textual Anonymisation [Internet]. In: Fifth International Conference on Language Resources and Evaluation. 2006 [cited 2023 2]. Available from: <https://aclanthology.org/L06-1110/>.
76. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. Journal of the American Medical Informatics Association [Internet] 2013 Jan. [cited 2023 2];20:84–94. Available from: <https://doi.org/10.1136/amiajnl-2012-001012>.
77. Lee JY, Dernoncourt F, Szolovits P. Transfer Learning for Named-Entity Recognition with Neural Networks [Internet]. In: Clinical Natural Language Processing. 2018 [cited 2023 2]. Available from: <https://aclanthology.org/L18-1708>.
78. Moqurrah SA, Ayub U, Anjum A, Asghar S, Srivastava G. An Accurate Deep Learning Model for Clinical Entity Recognition from Clinical Notes. Institute of Electrical and Electronics Engineers Journal of Biomedical and Health Informatics 2021;25. <https://doi.org/10.1109/JBHI.2021.3099755>
79. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Learning to identify Protected Health Information by integrating knowledge- and data-driven algorithms: A case study on psychiatric evaluation notes. J Biomed Inform [Internet] 2017 Nov. [cited 2023 2];75S:S28–33. <https://doi.org/10.1016/j.jbi.2017.06.005>
80. Young Lee J, Dernoncourt F, Uzuner O, Szolovits P, Albany S. Feature-Augmented Neural Networks for Patient Note De-identification [Internet]. In: Clinical Natural Language Processing. 2016 [cited 2023 2]. p. 17–22. Available from: <https://doi.org/10.1016/j.jbi.2017.06.005>.
81. Li XB, Qin J. Anonymizing and Sharing Medical Text Records. PMC- Author Manuscripts [Internet] 2017 Apr. [cited 2023 2];28:332–52. <https://doi.org/10.1287/isre.2016.0676>

