De-identification of Clinical Text for Secondary Use: Research Issues

Hanna Berg, Aron Henriksson, Uno Fors and Hercules Dalianis
Department of Computer and Systems Sciences, Stockholm University, Sweden
{*hanna.berg, aronhen, uno, hercules*}*@dsv.su.se*

Keywords: De-identification, Privacy, Electronic Health Records, Clinical Text, Natural Language Processing

Abstract: Privacy is challenged by both advances in AI-related technologies and recently introduced legal regulations. The problem of privacy has been extensively studied within the privacy community, but has largely focused on methods for protecting and assessing the privacy of structured data. Research aiming to protect the integrity of patients based on clinical text has primarily referred to US law and relied on automatically recognising predetermined, both direct and indirect, identifiers. This article discusses the various challenges concerning the re-use of unstructured clinical data, in particular in the form of clinical text, and focuses on ambiguous and vague terminology, how different legislation affects the requirements for de-identification, differences between methods for unstructured and structured data, the impact of approaches based on named entity recognition and replacing sensitive data with surrogates, as well as the lack of measures for usability and re-identification risk.

1 INTRODUCTION

Electronic health records (EHRs) are a valuable resource for research aimed at developing and evaluating health care, but they also contain information which may jeopardise personal integrity. Due to the sensitive nature of EHRs, access is restricted and protected by data protection laws, patient data laws or similar, which may for example require consent by patients and/or care organisations in order for such data to be used for research. However, if the risk of identification is deemed sufficiently low, patient consent may not be required. Automatic de-identification techniques aim to reduce the risk of identification and may therefore enable access to clinical data for research while protecting the privacy of patients.

In order to allow for secondary use of EHRs, in cases where informed consent is difficult to obtain, it is not sufficient to de-identify only structured data; one must also consider unstructured data. One type of unstructured data that needs to be considered is clinical text. Clinical text may describe, e.g., patient history, social background, relatives' contact information and the patient's living situation (Dalianis, 2018). Clinical text may, in fact, include more sensitive information than structured EHR data, but is unfortunately more challenging to de-identify. Research may also require access to both structured and unstructured EHR data, for example to study and develop new algorithms for decision support. There has been an increasing interest in AI-based solutions with advanced and data-hungry algorithms. The need for ever-increasing amounts of data has led to a need for well-functioning methods to protect patients' right to privacy, while maintaining high-quality data.

De-identification is the process of mitigating the risk of identifying individuals in datasets by altering the data. The most appropriate de-identification method and de-identification level may depend on the type of data, as well as the context in which the data will be used. There are methods for structured data to ensure that possibly identifying values are common enough, both separately and in combination, not to be identifying. With free text data, it is not as clear which information is stored in which section. Therefore, the primary method for de-identification is based on finding information belonging to certain pre-determined classes that are deemed to be possibly identifying. To that end, natural language processing and, specifically, named entity recognition (NER) is used, after which the identified information is obscured.

In this paper, a number of research issues concerning de-identification of unstructured clinical data – in the form of clinical text – are highlighted and discussed. The issues concern privacy regulations, terminology, de-identification methods and evaluation.

This paper is an extension of an abstract submitted to HealTAC 2020.

2 PRIVACY REGULATIONS

Privacy regulations and, specifically, data protection laws exist to protect individuals rights to protect personal data and privacy. As regulations vary across countries, so do requirements for de-identification.

In this article, two privacy regulations are described, HIPAA and GDPR; HIPAA is used in the United States and GDPR is used in the European Union. HIPAA is commonly used as the basis for de-identification (Stubbs et al., 2017, Meystre et al., 2010), and GDPR as it is relevant to the context in which we work. The HIPAA Privacy Rule was introduced in 2003, while GDPR was introduced in 2018. The American Health and Portability Act, HIPAA, regulates the privacy protection of Protected Health Information, PHI (HIPAA, 1996). PHI is essentially defined as all health information with individual identifiers. There are 18 PHI identifiers; if these are removed, the data is not considered to be sensitive. These identifiers include names, dates more specific than year, geographic data, contact information and any unique identifying number or other code. The method of removing PHI identifiers, named Safe Harbour, is one of two HIPAA-compliant de-identification methods. The other method, Expert Determination, instead requires that an expert applies mitigating methods based on statistics and mathematics until the risk of identification is very small.

The General Data Protection Regulation, GDPR, covers personal data, defined as data relating to an identified or identifiable natural person, which is a person who could be identified directly or indirectly (European Comission, 2016). Examples of identifiers are specific physiological or social attributes, which, when combined, point to one individual.

While HIPAA does not cover datasets where data of certain classes has been removed or the re-identification risk is very small, GDPR considers all data which could potentially be attributed to a natural person through supplementary information or a tool that could be used for re-identification of personal data. This includes other supplementary information only accessible by the data controller. A dataset which is not identifiable without supplementary information is pseudonymised. GDPR encourages the use of pseudonymisation as a privacy-protecting measure, but pseudonymised data is still considered sensitive. If there are no re-identification risks, the data is not considered as personal data and not covered by GDPR. The requirement of zero risk has been criticised as unattainable, with the claim that the only way to achieve this is by not disclosing any data at all (El Emam, 2013).

The different levels of risk required may amount to de-identification systems developed for HIPAA not being suitable for usage under GDPR. HIPAA and, specifically, Safe Harbour are often the basis for clinical text de-identification (Stubbs et al., 2017, Kushida et al., 2012, Meystre et al., 2010, Marimon et al., 2019). It is unknown whether these methods are appropriate under GDPR, or if there is a need for other methods with guarantees for privacy.

3 TERMINOLOGY

Another challenge is the lack of a consistent use of terminology within the research of *anonymisation* and *de-identification*. According to a literature review, the definition of both terms varies within the biomedical literature, and definitions are often vague or non-existent (Chevrier et al., 2019). Only around half of the articles provided a definition of the terms. Articles using both terms often made a distinction between the two terms where anonymisation is then most commonly refers to probabilistic and statistical techniques. According to the same review, the term de-identification more commonly refers to rule-based techniques where information belonging to predefined categories are removed. According to another literature review (Meystre et al., 2010), de-identification and anonymisation are often used interchangeably; however, de-identification means that explicit identifiers are obscured and that anonymisation implies that implicit identifiers cannot be used to identify individuals by linkage. HIPAA, on the other hand, uses the term de-identification regardless of whether the approach is rule-based, where specific predetermined information types are removed, or probabilistic, where statistical methods are applied to ensure that the risk of re-identification is sufficiently low. GDPR, on the other hand, uses the word anonymisation to refer to data with no re-identification risks, and pseudonymisation for data which is not identifiable in the absence of supplementing data. Here, the term de-identification will be used to refer to any type of method where personal information is hidden or obscured with the intention to protect the privacy of data subjects.

A dataset is sometimes referred to as de-identified when risk mitigation techniques have been applied to ensure a small enough re-identification risk. This term usage has been criticised as misleading since it implies that the dataset has a level of re-identification risk that is, in reality, not met (El Emam, 2013).

The term *pseudonymised* has more than one meaning. Pseudonymised may refer to the process of masking sensitive information with surrogates (Dalianis, 2019). Pseudonymisation may also refer to the usage of an alias that, with a key, may be linked to the real original data (European Comission, 2016). These are similar concepts, where the former could be seen as a version of the latter, but without keys. The distinction made in GDPR between pseudonymised data and anonymised data is that the pseudonymised data can be re-identified by linkage, while anonymised data cannot. This is similar to the distinction made by, for example, Meystre et al. (2010) concerning anonymization and de-identification.

The different uses of each term increase the risk for misunderstandings. A dataset can be described as de-identified, anonymised, pseudonymised and as personal data simultaneously by different people, depending on the definition for each term and the context. There needs to be an increased awareness surrounding the use of terminology in de-identification research, with explicit definitions.

4 DE-IDENTIFICATION METHODS

While the task of de-identification is difficult no matter the type of data, the process is especially difficult for unstructured data. The methods for risk mitigation and assessment used on well-structured data are specifically designed for structured data and therefore not readily applicable to noisy, unstructured data like free text. Figure 1 provides an architectural overview of the different approaches for preserving privacy of clinical data and specifically clinical text.

4.1 De-identification of Structured Data

There are, as previously mentioned, methods to statistically ensure that structured data is protected against identification. Some examples are: K-anonymity, l-diversity and differential privacy. K-anonymity ensures that multiple individuals share the same combination of identifying values for structured data (El Emam and Dankar, 2008). l-diversity instead ensures that there is enough diversity in sensitive values within a dataset (Machanavajjhala et al., 2007). Another method is differential privacy (Dwork et al., 2014), where aggregation and noise introduction is combined to create de-identified views of the data; each time a new data request is made, the level of noise is adjusted based on previous information given. Wagner and Eckhoff (2018) conducted a systematic review of different different privacy metrics for structured data, describing and discussing over 80 metrics.

While all of these methods may individually have their disadvantages and there are risks associated with not doing them properly, they do offer the possibility of assessing the level of risk of re-identification, and then mitigating the risk. Differential privacy may also provide statistical guarantees against what could be inferred from the information provided (Dwork et al., 2014), while for example k-anonymity does not (Machanavajjhala et al., 2007).

There are a few examples of structured data and methods being used to de-identify unstructured data in the form of text, for example applying k-anonymity on quasi-identifiers identified with NER (Gardner and Xiong, 2009). Another example is recursive partitioning to cluster medical text records based on information similarity and value-enumeration to de-identify potentially identifying information (Li and Qin, 2017). These are promising alternatives that may provide additional flexibility and ways to deal with potentially identifying quasi-identifiers.

If a machine learning model is built using sensitive data, the process can be re-engineered and the individuals revealed. Papernot et al. (2017) have proposed a solution to this by injecting noise in the trained model by using differential privacy. In Figure 1, one can observe that adding external databases may enable the re-identification of individuals, so called data linkage. Encryption of data and the use of synthetic data are other methods for protecting privacy.

4.2 De-identification of Clinical Text

Systems intending to de-identify text generally rely on NER (Meystre et al., 2010), which is the task of locating and classifying named entities in unstructured text (Nadeau and Sekine, 2007). Using NER has the potential of being useful within the scope of HIPAA to enable data sharing. The strategy is to, through hand-crafted rules and/or machine learning, find entities belonging to any of the HIPAA classes, possibly with the addition of other classes which are deemed identifying (Stubbs et al., 2017). If any of these are found, they are either marked as belonging to a certain class, or replaced with similar data of the same class. A common method is to combine machine learning for the irregular or less structured entities, and rules for the more structured or regular entities within the free text (Meystre et al., 2010).

4.2.1 Identifying Sensitive Information

The early de-identification systems, for example the Scrub system (Sweeney, 1996), are rule-based. Rule-based approaches rely on rules, patterns and gazetteer lists. Since they rely on hand-crafted rules, little or

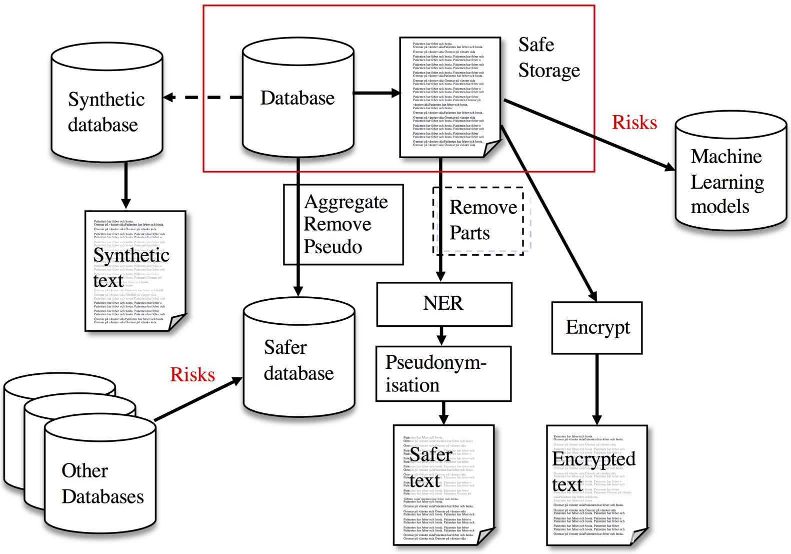

Figure 1: An architectural overview of the different approaches to preserve privacy in clinical data. The protected data is shown in the red rectangle. For structured data, methods like aggregation, generalization, pseudonymisation and noise perturbation may be used to create a safer database. For unstructured text, sections known to have a high density of sensitive information and low relevance for the task may be removed before named entity recognition is performed to find sensitive information. The identified sensitive information is then replaced with, e.g., pseudonyms to make the text safer. Beyond this, there are methods for creating synthetic databases and privacy preserving learning techniques for machine learning models

no annotated data is needed other than for evaluation purposes (Meystre et al., 2010). Crafting rules is, however, a complex task where the developers need to be aware of all possible PHI patterns that can occur. They also typically require customisation to a particular dataset and are therefore less generalisable. Supervised machine learning methods do not require hand-crafted rules, but require annotated data for the algorithm to train on. Feature-based supervised learning approaches rely on feature engineering. Common features are lexical features (e.g. word casing, word shape, punctuation, numerical characters), syntactic features (e.g. part-of-speech tags) and semantic features (e.g. terms from dictionaries, semantic types). Section headers may also be used. Unlike rule-based methods, supervised machine learning methods can automatically learn to recognise complex patterns. For telephone numbers or other data that tends to be regular and where the patterns are not complex, rule-based methods are still often used within hybrid systems in which rules and machine learning methods are combined. The first neural network for de-identification was introduced in 2016 (Dernoncourt et al., 2017). Neural networks can effectively learn features through composition over token embeddings and therefore do not require handcrafted features or feature engineering to the same extent as feature-based machine learning methods. These embeddings can be initialised randomly or pre-trained on large unlabeled data sets.

A number of shared task challenges have been organised to drive the development of de-identification systems forward. The challenges have focused on the identification of personally identifying information. As a part of the i2b2 project, three challenges have been organised: the first in 2006 (Uzuner et al., 2007), the second in 2014 (Stubbs et al., 2015) and the most recent in 2016 (Stubbs et al., 2017). Furthermore, a de-identification challenge on Spanish synthetic health records, MEDDOCAN, was organized during IberLEF 2019 (Marimon et al., 2019). During the first challenge in 2006, the submitted systems were either supervised feature-based machine learning systems, rule-based systems or a combination of the two (Uzuner et al., 2007). Machine learning method such as SVMs, CRFs, hierarchial Hidden Markov Models and decision trees were used. The best performing systems were CRFs or decision trees with rule template features. The pure rule-based systems performed the worst. The system with the best performance scored an entity-based binary precision score of 0.99 and a recall score of 0.98 on classes based on HIPAA's Safe Harbour (Stubbs et al., 2015). In the 2014 i2b2 challenge, a majority of the submitted systems were hybrid systems using CRFs and rules, and these also performed the best. The system with the best performance scored an entity-based binary precision score of 0.99 and a recall score of 0.96

on classes based on HIPAA's Safe Harbour. In the 2016 i2b2 de-identification task on psychiatric intake records, CRFs were still the most popular approach to the de-identification task, but among the top five, two teams also used LSTMs (Stubbs et al., 2017). While the most common approach was supervised machine learning, three of the top four systems were hybrid systems in which both multiple machine learning techniques were used but also hand-crafted rules. Similarly, during the MEDDOCAN anonymisation challenge with synthetic Spanish electronic health records narrative text, deep learning systems outperformed other systems (Marimon et al., 2019). The best performing system was a bidirectional LSTM with FLAIR embeddings, as well as both domain-independent and domain-dependent fastText embeddings (Lange et al., 2019). This system achieved an entity binary precision score of 0.98 and a recall score 0.97 (Marimon et al., 2019).

NER is used to find direct and indirect identifiers in order to remove them. The removal of, e.g., names, contact information and serial codes may protect against re-identification. Other types of information are, however, difficult to handle using only NER methods. Diagnosis codes, and the diagnoses themselves, have been shown to be identifying. According to Loukides et al. (2010), 96% of all individuals in 2,600 patient records could be identified through their diagnosis codes. While unique combinations of diseases along with years may appear in written form within a patient record, it is not considered to be an identifier according to HIPAA, and is not something that traditional de-identification of text can handle. It is important to be aware of these limitations, and to seek ways to overcome them.

De-identification of text is so heavily connected to the NER step that the main evaluation metrics of text de-identification are the ones used for NER: recall, precision and F_1-score (Meystre et al., 2014a). These metrics measure how many entities of each predefined class a system manages to find and classify correctly, but does not consider how identifying the kept in formation is. In reality, a first name could be either very common, as for example *John Smith* and not identifiable, but also unique, as for example *Severus Snape*. Similarly, a disclosed phone number to a close relative increases the risk of identifying a patient to a greater degree than the number to a hospital. Similarly, the incorrect labelling of an entity as a PHI may affect downstream tasks to different degrees. The labelling of *Parkinsons* as a surname could be assumed to potentially cause trouble, and the incorrect identification of *the health care unit* as a specific health care unit, is likely less harmful.

4.2.2 Obscuring Identifiers

If a token is classified as belonging to a possibly identifying class, it is either removed or replaced. The removal may either be in the form of masking, or keeping information about which type of information is identified. The method of instead replacing data with similar data has both its advantages and disadvantages.

Replacing a name with a surrogate may both lead to an increase or decrease of re-identification risk. A system relying on NER will unlikely be able to find all possible sensitive data. The use of realistic surrogates, as opposed to masking, conceals information about which data is real and which data is not (Carrell et al., 2012). This is called Hidden In Plain Sight, HIPS.

The use of realistic surrogates, or generalising the information, may at the same time allow for some information to be kept. The total removal of temporal information may make the dataset insufficient for some research, whereas the use of surrogates would maintain the usability of the dataset. There is, at the same time, a risk of altering clinical information which should be kept as it is, with a possible increase in the risk of false conclusions (Meystre et al., 2014a).

5 EVALUATION

There are two aspects of de-identification methods that are evaluated: the risk of re-identification and the impact de-identification has on downstream tasks.

5.1 Re-identification

Re-identification is the identification of individuals in a dataset which is claimed to be de-identified. This is most commonly achieved by linkage of data between data sources. The desire to use interconnected medical data for various sources may lead to a higher risk of re-identification through linkability.

Most, if not all, examples of automatic re-identification include structured data. This may have various causes: firstly, it is a more common data source; secondly, there are strategies for assessing the risk of re-identification; finally, linking information from various data tables is not a complex task. There are examples where information in newspaper text has been extracted manually and then linked to de-identified structured data to successfully re-identify individuals (Yoo et al., 2018); however, there are few examples of re-identification from unstructured text.

Re-identification studies on de-identified text have focused on the risk of re-identification by another individual. Carrell et al. (2012) showed that by replacing identified sensitive data with surrogates, the risk of humans finding residual identifiers not found by the NER system significantly decreased. Meystre et al. (2014b) set up an experiment where physicians were asked if they could recognise their patients in a dataset. In 4.65% of the 86 pseudonymised discharge summaries, the physician thought they recognised a patient, but in no cases a patient was correctly identified. A study by Grouin et al. (2015) showed that, for de-identified text, it was possible to recover specific values. The disclosed information was, however, not sufficient to re-identify any patient unless the adversary had access to the hospital health information system and several documents from the same patient.

There are no clear estimates on how big the risk of re-identification is with the Safe Harbour method on unstructured data. In general, there are few methods for assessing the risk of re-identification for unstructured data. The most common metric relating to the safety of a de-identification system for text is based on calculating how many sensitive entities are found in a test set. While this recall measurement is likely to correlate with the risk of re-identification, it does not directly measure this.

Since there are no examples known to us of automatic re-identification of text in the same way as for structured data, it is difficult to determine how such an attack would be performed and therefore also difficult to determine the level of effort required. In the end, this makes it difficult to determine the needs for de-identification in practice. From one perspective, a step toward improving de-identification would be to investigate possible re-identification techniques.

5.2 Impact of De-identification

There is a concern that de-identification would affect data quality. It has been hypothesised that the de-identification process may be harmful on downstream tasks as clinical information erroneously classified as PHI may lead to a reduction in information content and the introduction of misleading information (Meystre et al., 2014a). It has also, however, been hypothesised that de-identification may potentially improve machine learning performance by reducing dimensionality and noise (Obeid et al., 2019).

Studies have so far shown no significant differences between using original text data or de-identified text data as training data for text classification (Obeid et al., 2019), and small but possibly statistically significant benefits of training on de-identified data for medication name extraction (Deleger et al., 2013). Meystre et al. (2014a) noted that between 1.2-3% fewer SNOMED-CT concepts were found in the de-identified dataset than in the original version, but the difference was largely explained by PHIs being erroneously recognised as SNOMED-CT concepts in the original version rather than a decrease in information content. So far, no significant deterioration has been documented for de-identification systems on downstream tasks. It could, however, be assumed that a de-identification system with too low precision reduces the amount of available information in a way that negatively affects downstream tasks (Berg et al., 2020).

Uncertainty about the impact de-identification has on the dataset results in uncertainty about which measures to use in order to compare systems. In practice, F_1-score, a weighted score between precision and recall, is often used to compare systems (Stubbs et al., 2017, Meystre et al., 2010). Studies point to the fact that the precision of de-identification systems does not necessarily have a significant impact on the content of other research (Obeid et al., 2019, Meystre et al., 2014a), perhaps because there are so few sensitive tokens in total. There is, however, a need to determine what impact precision has, in order to be able to determine what weight should be placed on this when comparing and evaluating systems – but also to decide how they should be built.

A case where de-identification, however, seems to have an impact on the end product is when using de-identified data for training a NER de-identification system (Berg et al., 2019, Yeniterzi et al., 2010). According to these studies, models trained on datasets with surrogates will perform worse on real data than a model trained on real data would.

6 OTHER APPROACHES

There are other privacy-protecting approaches, including synthesised datasets, encryption of data and black box tools.

Synthesised data is data that is not real but has real properties. Synthetic data can be created manually, such as in Rama et al. (2018), or be multiply-imputed synthetic microdata with exactly the same statistical information as the real microdata and should lead to the same statistical inferences (Nowok et al., 2016). Methods, such as Synthpop, for generating synthetic data with the same statistical information rely on and generate only structured data.

In one approach by Dalianis and Boström (2012), all the lexical items themselves were removed, while word features, such as part-of-speech, were kept and used for the machine learning, resulting in only a slightly decreased performance. For structured data, complete encryption has been carried out before machine learning algorithms have been applied (Bos et al., 2014, Arellano et al., 2018).

Finally, an alternative approach for research relying on machine learning is that the software is developed and possibly trained externally, training on non-sensitive data, in relation to where the sensitive clinical text is stored, and then the models are applied or executed by a gate-keeper internally and reported to the researcher (Almgren et al., 2016). This is a viable alternative for research related to clinical text mining.

7 CONCLUSIONS

We have discussed various research issues regarding the preservation of patients' integrity in the reuse of clinical text. The main method is based on HIPAA and Safe Harbour. This involves first determining what kind of information is potentially identifying, finding entities that belong to these classes, and then deleting or obscuring them. There are no methods to statistically calculate which information is identifying, but the method relies instead on rules. While this may be appropriate in a US context, it is unclear whether this type of de-identification allows for the sharing of data to researchers in a European context since the risk of re-identification remains unclear. There is a need for other methods to determine which information is sensitive and managing this, possibly by taking advantage of available structured data. At present, de-identification through named entity recognition seems, however, to be the best we have.

REFERENCES

Almgren, S., Pavlov, S., and Mogren, O. (2016). Named Entity Recognition in Swedish Health Records with Character-Based Deep Bidirectional LSTMs. *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), held in conjunction with Coling 2016*, pages 30–29.

Arellano, A. M., Dai, W., Wang, S., Jiang, X., and Ohno-Machado, L. (2018). Privacy policy and technology in biomedical data science. *Annual review of biomedical data science*, 1:115–129.

Berg, H., Chomutare, T., and Dalianis, H. (2019). Building a de-identification system for real swedish clinical text using pseudonymised clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 118–125.

Berg, H., Henriksson, A., and Dalianis, H. (2020). The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *To appear in the Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (LOUHI 2020).*

Bos, J. W., Lauter, K., and Naehrig, M. (2014). Private predictive analysis on encrypted medical data. *Journal of biomedical informatics*, 50:234–243.

Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., and Hirschman, L. (2012). Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., and Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *Journal of medical Internet research*, 21(5):e13484.

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer.

Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Dalianis, H. and Boström, H. (2012). Releasing a Swedish Clinical Corpus after Removing all Words – De-identification Experiments with Conditional Random Fields and Random Forests. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC*, pages 45–48.

Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L., and Solti, I. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

El Emam, K. (2013). *Guide to the de-identification of personal health information*. Auerbach Publications.

El Emam, K. and Dankar, F. K. (2008). Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637.

European Comission (2016). Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

Gardner, J. and Xiong, L. (2009). An integrated framework for de-identifying unstructured medical data. *Data & Knowledge Engineering*, 68(12):1441–1451.

Grouin, C., Griffon, N., and Névéol, A. (2015). Is it possible to recover personal health information from an automatically de-identified corpus of french ehrs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39.

HIPAA (1996). Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed: 2020-01-17.

Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., and Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82.

Lange, L., Adel, H., and Strötgen, J. (2019). NLNDE: The Neither-Language-Nor-Domain-Experts' Way of Spanish Medical Document De-Identification. *arXiv preprint arXiv:2007.01030.*

Li, X.-B. and Qin, J. (2017). Anonymizing and sharing medical text records. *Information Systems Research*, 28(2):332–352.

Loukides, G., Denny, J. C., and Malin, B. (2010). The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327.

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

Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodrguez, H., Lopez Martin, J., Villegas, M., and Krallinger, M. (2019). Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS. org), Bilbao, Spain (Sep 2019), TBA.*

Meystre, S. M., Ferrández, Ó., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2014a). Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of biomedical informatics*, 50:142–150.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70.

Meystre, S. M., Shen, S., Hofmann, D., and Gundlapalli, A. V. (2014b). Can physicians recognize their own patients in de-identified notes? In *MIE*, pages 778–782.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11):1–26.

Obeid, J. S., Heider, P. M., Weeda, E. R., Matuskowitz, A. J., Carr, C. M., Gagnon, K., Crawford, T., and Meystre, S. M. (2019). Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in health technology and informatics*, 264:283.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. *Proccedings of 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017.*

Rama, T., Brekke, P., Nytrø, Ø., and Øvrelid, L. (2018). Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 111–121.

Stubbs, A., Filannino, M., and Uzuner, Ö. (2017). De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18.

Stubbs, A., Kotfila, C., and Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Wagner, I. and Eckhoff, D. (2018). Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38.

Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Hirschman, L., and Malin, B. (2010). Effects of personal identifier resynthesis on clinical text de-identification. *Journal of the American Medical Informatics Association*, 17(2):159–168.

Yoo, J. S., Thaler, A., Sweeney, L., and Zang, J. (2018). Risks to patient privacy: A re-identification of patients in maine and vermont statewide hospital data. *Technology Science. Oct.*