INCOGNITUS: A Toolbox for Automated Clinical Notes Anonymization

Bruno Ribeiro
Fraunhofer Portugal AICOS
R. Alfredo Allen 455,
4200-135 Porto

Vitor Rolla
Fraunhofer Portugal AICOS
R. Alfredo Allen 455,
4200-135 Porto

Ricardo Santos
Fraunhofer Portugal AICOS
& LIBPhys, Physics Department
FCT/NOVA, 2829-516 Caparica

Abstract

Automated text anonymization is a classical problem in Natural Language Processing (NLP). The topic has evolved immensely throughout the years, with the first list-search and rule-based solutions evolving to statistical modeling approaches and later to advanced systems that rely on powerful state-of-the-art language models. Even so, these solutions fail to be widely implemented in the most privacy-demanding areas of activity, such as healthcare; none of them is perfect, and most can not guarantee rigorous anonymization. This paper presents INCOGNITUS, a flexible platform for the automated anonymization of clinical notes that offers the possibility of applying different techniques. The available tools include an underexplored yet promising method that guarantees 100% recall by replacing each word with a semantically identical one. In addition, the presented framework incorporates a performance evaluation module to compute a novel metric for information loss assessment in real-time.

1 Introduction

The digitization of medical textual data has allowed for a whole new range of possibilities, such as the development of tools for the summarization of clinical notes or the automated identification of the International Classification of Diseases (ICD) codes in clinical text. However, clinical data contains sensitive information regarding both patients and health professionals. These entities are protected by the General Data Protection Regulation (GDPR), which grants equal data protection rights to all European Union (EU) citizens (GDPR, 2018). Therefore, information systems must be compliant to maintain such information private. This poses a challenge when considering the publication of clinical data for secondary usage.

The GDPR (GDPR, 2018) defines anonymization as the process through which "personal data (is) rendered anonymous in such a manner that the data subject is not or no longer identifiable" (Recital 26). However, achieving true anonymization is not a simple task, especially when considering unstructured data such as clinical notes. In fact, despite the fact that many literature works propose strategies for the automated anonymization of clinical text, their implementation in real contexts is still scarce. Consequently, the problematic access to clinical text data for secondary usage remains a barrier to scientific research.

This demonstration paper presents the INCOGNITUS platform for automated clinical notes anonymization. The beta version is available online at `https://tospe-incognitusfhp.streamlit.app/`. An overview video can be watched at `https://www.youtube.com/watch?v=4lePn19ZwJE`. Besides offering a wide range of methods to perform anonymization tasks, INCOGNITUS was designed to address two common flaws of similar systems: (i) the inability to provide truly anonymized outputs (100% sensitive information removal) (Abdalla et al., 2020) and (ii) the lack of an assessment on the quantity of relevant information that gets lost in the anonymization process (Mozes and Kleinberg, 2021).

The remainder of the manuscript is organized as follows. The next section contextualizes text anonymization literature, mainly focusing on the clinical domain. Then, the platform proposed is described, along with its composing methods and models. The experimental setup followed to develop each component is detailed in Section 4. The results associated with these experiments are then discussed in Section 5. In the last section, the main conclusions of this work are presented.

2 Background

The list of literature publications focused on clinical text anonymization is extensive. This section presents a chronological overview of the developments achieved in this area of research, culminating

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics
System Demonstrations, pages 187–194

in a final discussion on the main findings regarding the strengths and flaws of current techniques and future research tendencies.

2.1 First Approaches

The first works on clinical text de-identification involved the development of simple systems relying on handcrafted sets of rules, dictionary and medical thesaurus lookups, and pattern matching algorithms (Sweeney, 1996; Ruch et al., 2000; Thomas et al., 2002; Gupta et al., 2004; Douglass et al., 2004). These methods established the potential of applying Natural Language Processing (NLP) techniques to the anonymization problem, reporting performances between 94% and 99% in terms of recall. Even so, since they were usually highly adapted to the characteristics shared by the very notes that they were tested upon, their generalization ability was poor.

2.2 Conditional Random Fields

Following these early developments, two events boosted the scientific knowledge around clinical text data de-identification: the 2006 and 2014 Informatics for Integrating Biology to the Bedside (i2b2) challenges (Uzuner et al., 2007; Stubbs et al., 2015). These competitions encouraged researchers to produce innovative solutions to approach unstructured data anonymization. At this point, solutions based in Conditional Random Fields (CRF) took over as the best-performing methods (Aramaki et al., 2006; Wellner et al., 2007; Gardner and Xiong, 2008; Dehghan et al., 2015; Liu et al., 2015; Yang and Garibaldi, 2015). These systems leveraged features such as Part-of-Speech (POS) tags, surrounding words, position within the document, word form, and capitalization to identify sensitive information within the text. Some of them also included regular expression matching and list search modules along with the main CRF model (Wellner et al., 2007; Dehghan et al., 2015; Liu et al., 2015; Yang and Garibaldi, 2015). In 2006, all the proposed methodologies achieved f1-scores higher than 95%, while the best-performing method in 2014 reported a value of 93.6%.

2.3 Deep Learning

As the scientific community's focus shifted towards Deep Learning (DL) strategies, many studies were published where these models were trained on the anonymization task. Most of these studies implement long short-term memory (LSTM) recurrent neural networks (RNN), which are known to be effective in Named Entity Recognition (NER) tasks (Dernoncourt et al., 2016; Liu et al., 2017; Stubbs et al., 2017; Yang et al., 2019). Some of these consisted of hybrid approaches, complementing the DL models with CRF-based and even simpler modules.

Friedrich et al. (Friedrich et al., 2019) proposed an adversarial learning approach based on an LSTM-CRF architecture. Their solution prevents two procedures that can be used to re-construct Personal Health Identifiers (PHIs): (i) the development of a model which can learn the reverse transformation mechanism; and (ii) a look-up table with all the inputs and their respective representations. An adversarial representation is trained to perform two tasks directly following these two procedures. The goal is to achieve a secure solution where the best adversarial/negative representation cannot re-construct the PHIs.

Recently, Abdalla et al. (Abdalla et al., 2020) presented an innovative approach that leverages proximity measures between word embeddings. They argue that solutions based on NER techniques are insufficient to guarantee that no sensitive information gets overlooked, as search algorithms are never perfect. To counter this possibility, they propose the substitution of each token with a semantically proximate from the embedding space. This obfuscation strategy ensures that all sensitive information gets removed at the cost of a low readability. Even so, the authors report that little influence is observed on clinical machine learning tasks by taking the anonymized data as input (up to a 5% decrease in f1-score) since the contextualized token substitution allows for the preservation of data patterns that are useful in those tasks.

2.4 Discussion

While most of the discussed methods present extremely high performance considering the traditional metrics (f1-score and recall), their success remains mostly scientific, and thus the problem of automated text anonymization remains unsolved. This might be due to problems such as low generalization to external data and high production of false positives, which constitute huge barriers to real-world deployment and application. The Track 1A of the CEGS de-identification challenges (Stubbs et al., 2017) exposed these limitations by using test data collected from a different source than the train data. Although this is a fundamental measure

to attain a reliable performance assessment, it is not usually adopted in most literary works around this topic (Yang et al., 2019). The state-of-the-art methods could not maintain their success rates in the mentioned task. The higher-performing solution presented an f1-score of approximately 80%, way below the above-90% scores reported in most literary works.

A recent paper by Mozes and Kleinberg (Mozes and Kleinberg, 2021) alerted for the necessity of reliable performance assessment strategies, putting forward an innovative method based on three criteria: i) technical evaluation through the commonly used evaluation metrics, ii) information loss estimation, and iii) de-anonymization tests. In addition, Pilán et al. (Pilán et al., 2022) also proposed a new set of metrics focused on privacy protection and utility preservation to evaluate text-anonymization solutions.

3 INCOGNITUS Framework

INCOGNITUS is a flexible and intuitive toolbox aimed at transparent and accountable automated text anonymization. This framework offers the possibility to employ different techniques while providing performance measures relating to the anonymization task itself and the resultant loss of information. All this content is displayed and accessible through an interactive, user-friendly interface, which can be seen in Figure 1. With these insights, the user can consciously select the adequate approach to anonymization, considering the specifications of each application context.

The flow of information circulating through INCOGNITUS is represented in the diagram of Figure 2. First, the user uploads or writes the content of a clinical note. The anonymization may then be performed through any of the available techniques. As the anonymization is completed, an ICD-10 classification model runs in the background, and its outputs are leveraged to estimate the ratio of information that got lost in the anonymization process. By the end of the calculation, this measure is displayed in the user interface, as well as the values of the standard performance metrics (recall, precision, and f1-score) associated with the technique selected. The user can then choose to (i) download the anonymized content or (ii) select another technique and repeat the process.

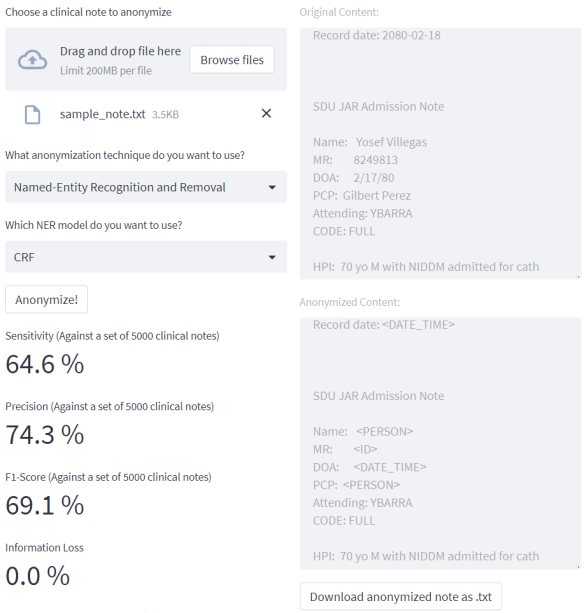

Figure 1: INCOGNITUS interface's outlook. In the depicted scenario, the user selected a NER approach based on a CRF model. On the right side of the page, the user can consult the contents of the original (top) and anonymized (bottom) versions of the uploaded sample note. A performance report is presented on the bottom left corner of the page, regarding the estimated values of recall, f1-score, precision, and Information Loss for the selected approach.

3.1 Components

Currently, INCOGNITUS comprises four NLP models that fulfill different purposes. First, there are two NER models for de-identification: a CRF classifier (Lafferty et al., 2001) and a pre-trained Spacy model (Montani et al., 2020). A Word2Vec embeddings model (Mikolov et al., 2013) is used to fuel the K-Nearest Embeddings Obfuscation technique (KNEO) based on the work of Abdalla et al. (Abdalla et al., 2020). And finally, a pre-trained BioBERT-based model (Lee et al., 2020) fine-tuned on the MIMIC III dataset (Johnson et al., 2016) is employed to infer the information loss over an ICD-10 classification problem.

3.1.1 Named Entity Recognition and Removal

CRF classifiers are standard in NLP and considered reliable benchmarks in de-identification tasks. On the other hand, Microsoft Presidio (Mendels and Balter) is a powerful tool designed to ensure that sensitive data, such as credit card numbers, names, locations, and financial data, is appropriately identified and anonymized in text. It comprises two modules: an analyzer based on NER techniques to

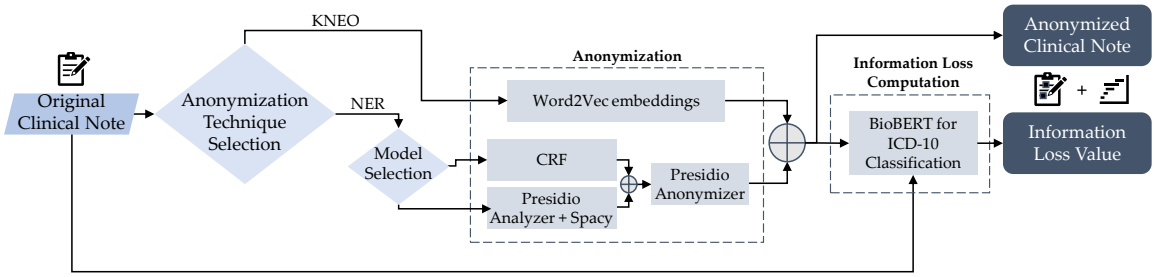

Figure 2: INCOGNITUS pipeline flowchart. After the clinical text content is uploaded to the platform, the user selects the technique and model desired to perform the anonymization task. The anonymization is performed, by employing the methodologies selected. Both the output of the anonymization phase and the original content is then fed into an ICD-10 classification model, which provides lists of the top ICD-10 code categories associated with each input. These results are leveraged to estimate the loss of information associated with the anonymization process. Finally, the user interface presents the anonymized content and the information loss estimate, along with other traditional metrics measured a priori. At the end of this process, the user may choose to download the anonymized version of the uploaded content or repeat the whole process using a different anonymization technique.

determine the sensitive entities and an anonymizer that takes the locations of those entities and removes or replaces them.

INCOGNITUS offers two possibilities for the recognition of sensitive entities: a CRF model trained upon the joint content of the training datasets of the i2b2 challenges (Uzuner et al., 2007; Stubbs et al., 2015), and an implementation of Presidio's analyzer module, receiving a Spacy model as input. Spacy is a Python package that provides accurate and up-to-date language models.

Once the identification of entities is performed through either of these NER-based techniques, the outputs are provided to the Presidio anonymizer, which returns a version of the original note where categorized tags replaced the identified entities.

3.1.2 K-Nearest Embeddings Obfuscation

A fundamental component of INCOGNITUS is its obfuscation module, which integrates a version of the algorithm proposed by Abdalla et al. (Abdalla et al., 2020). This approach is referred to as K-Nearest Embeddings Obfuscation (KNEO) in this work. It consists of randomly replacing each token composing a given text with one of the K semantically nearest ones within a space of embeddings. This methodology guarantees 100% recall in the anonymization task since no token is left unaltered.

3.1.3 Performance and Information Loss

Another innovative feature of this toolbox is the provision of a case-specific estimation of the information lost during the anonymization process, apart from the general performance metrics associated with each solution provided.

Before and after the anonymization of a clinical record, a pre-trained model (BioBERT fine-tuned on MIMIC III data (Lee et al., 2020; Johnson et al., 2016)) is used to identify the top N most frequent ICD-10 code categories (out of 157 possible) of both the original and anonymized versions of the same document. The value of Information Loss (IL) is inferred by dividing the number of classes simultaneously present in both prediction lists by the number N of top categories considered. Equation 1 presents the formalization of the proposed metric, where y_{orig} and y_{anon} represent the list of top N codes predicted in the original and anonymized versions of the same document, respectively. In INCOGNITUS, $N = 10$.

$$IL = (1 - \frac{\sum_{i=1}^{N} (y_{anon_i} \in y_{orig})}{N}) \times 100 \quad (1)$$

4 Experimental Setup

4.1 Datasets

Three distinct datasets were used to develop and evaluate the different anonymization strategies available at the INCOGNITUS platform.

MIMIC-III (Johnson et al., 2016) is an extensive, freely-available database comprising health-related data of over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. This dataset contains 1.2 million notes, including radiology reports, nursing notes, and discharge summaries.

The i2b2 project organized two challenges on automatically removing PHIs from medical records

(Uzuner et al., 2007; Stubbs et al., 2015). Both datasets released for the competitions are utilized in this research. The 2006 dataset contains 669 notes for training and 220 for testing, while the one used in 2014 counts 790 training samples and 514 testing samples.

4.2 Preprocessing

The original splits of both i2b2 datasets were maintained. As for the MIMIC III data, only discharge summaries were considered for this study, totaling 59,652 clinical notes, of which 5,000 were selected to test the different anonymization approaches. The remaining 54,652 were used to train the Word2Vec embeddings model employed in the KNEO approach.

Both i2b2 datasets contain fake PHI to simulate real, non-anonymized clinical records. The same is not valid for the MIMIC III notes, where all the sensitive information appears replaced by categorized entity tags. In order to re-establish a realistic clinical note structure, the Faker library for Python (Faraglia) was used to create fake entities according to each tag category.

4.3 NER Models

The first method developed was the CRF model. For this purpose, the following features were considered regarding each token and their immediate neighbors: POS tag, the last 2 and 3 constituting characters, whether it starts with a capital letter, whether it is a title, and whether it is a digit. This model was trained upon the train datasets of both i2b2 challenges, adding up to 1389 clinical notes.

For the second NER approach, Presidio Analyzer was used together with a language model from spacy, pre-trained on the OntoNotes 5 dataset (Ralph Weischedel, 2013), and reporting F1-score and recall values of 90% on the NER task. This solution was used as an off-the-shelf method; therefore, the language models were not re-trained upon any of the working train sets.

4.4 Word Embeddings

The training of embedding models is a complex task that requires extensive data. As such, the INCOGNITUS embeddings model was trained upon the subset of 54,652 discharge summaries retrieved exclusively from MIMIC III, following a Word2Vec strategy (Mikolov et al., 2013).

5 Results

To estimate the performance of the offered strategies in the anonymization task, they were tested against the working test sets. For each experiment, the f1-score regarding the binary task of classifying sensitive information was calculated. The mean IL associated with anonymizing each clinical note was also measured, based on Equation 1. The results of this experiment are displayed in Table 1.

		f1-score	**IL**
$i2b2_{2006}$	CRF	94.8	15.8 ± 11.4
	Presidio	73.0	21.6 ± 13.0
	KNEO	-	$59,9 \pm 21.3$
$i2b2_{2014}$	CRF	87.8	15.7 ± 12.4
	Presidio	64.6	21.3 ± 14.0
	KNEO	-	58.4 ± 21.1
$MIMIC$	CRF	69.1	21.3 ± 13.8
	Presidio	66.6	24.9 ± 14.6
	KNEO	-	63.4 ± 18.4

Table 1: Values of f1-score and mean $\pm$ standard deviation of the IL, attained by each of the anonymization methods offered by INCOGNITUS, in each test set.

By looking at the f1-score values in Table 1, it is clear that achieving high performance through fairly simple NER techniques is possible, as is evident by the results attained by the CRF model. These are in line with most reported values in the literature for similar tasks, when the data used for the test follows the same structure as the training data. Suppose one considers the results attained through the Presidio Analyzer instead, which uses a model trained upon the OntoNotes dataset. In that case, one can observe a clear drop in performance compared to that achieved with CRF. This might seem strange initially, considering that the configuration used by Presidio incorporates much more complex methods than those involved in training the CRF model. The fact that the NER models used by Presidio were not re-trained in any of the available clinical datasets might be the reason behind this. In fact, when the CRF model was tested against an external dataset (MIMIC III discharge notes subset), it presented a drop in performance of almost 20%, much closer to the values attained by presidio. These results are concerning because expecting access to notes for training in real application contexts is somewhat unrealistic. To have that, one would need access to content anonymized in a non-automated way or not anonymized at all,

which defeats the purpose of these tools.

Furthermore, while f1-scores of 95% are high, these values can never comply with the GDPR definition of anonymization. Even if these rose to 100%, the underlying risk of existing particular identifiers not being considered during training is still a threat. In this regard, the traditional methods of automated anonymization (NER-based) cannot compete with the KNEO strategy, since the latter replaces every single token from the original text, guaranteeing no such occurrences. Since every original token is exchanged, the f1-score calculation becomes inappropriate for the KNEO methodology. On the one hand, by replacing every token, it is guaranteed that none of the original content gets overlooked, ensuring 100% recall. On the other hand, since the replacement of every original word is the fundamental ideology behind this method, the concept of false positives makes little sense. One could argue that every non-sensitive token replaced might be regarded as a false positive. Still, such an interpretation ignores the role of neighbor word embeddings in preserving the information encoded.

Even so, although KNEO outperforms both NER-based strategies in preventing sensitive information leaks, the quantity of relevant information lost in the process is undeniably higher than what is observed for the other methods. Looking at the information loss values, this tendency is clear: around 60% of the original content is lost when applying the KNEO strategy. This means that, on average, the obfuscator hides the information related to six of the ten codes identified in the original notes. This becomes even more concerning if we consider that the classification task used to compute the IL metric simplifies a much more complex one: identifying individual ICD10 codes. Therefore, applying this strategy before performing advanced text processing tasks, such as clinical note summarization, could be problematic. The rates of information lost for the NER-based strategies are much lower, although it is visible that some information also gets hidden (around 20%).

In sum, since no technique is flawless, it is fundamental to understand (i) the context of application and (ii) the pros and cons of applying each technique. INCOGNITUS answers to the second necessity by providing various solutions and estimating performance during anonymization. In this way, it allows the users to switch between strategies according to their needs and the characteristics of each method.

6 Conclusions

This paper introduced INCOGNITUS, a user-friendly platform to prompt conscious automated anonymization of clinical text. It provides two distinct NER-based methods and a recently proposed alternative that guarantees proper anonymization (100% recall) at the cost of information loss and low readability. Employing a pre-trained classifier of ICD-10 code categories, INCOGNITUS brings a new way of estimating the information lost during anonymization. This framework moves the research around clinical text anonymization forward, towards accountable strategies and fair performance assessments.

Acknowledgements

This article is a result of work conducted under two projects: "ConnectedHealth" (n.°46858), supported by the Competitiveness and Internationalisation Operational Programme, Portugal (POCI) and Lisbon Regional Operational Programme (LISBOA 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF); and "CardioFollow.AI: An intelligent system to improve patients' safety and remote surveillance in follow-up for cardiothoracic surgery", and supported by national funds through 'FCT – Portuguese Foundation for Science and Technology, I.P.', with reference DSAIPA/AI/0094/2020.

References

Mohamed Abdalla, Moustafa Abdalla, Frank Rudzicz, and Graeme Hirst. 2020. Using word embeddings to improve the privacy of clinical notes. *Journal of the American Medical Informatics Association*, 27(6):901–907.

Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. 2006. Automatic deidentification by using sentence features and label consistency. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Azad Dehghan, Aleksandar Kovacevic, George Karystianis, John A. Keane, and Goran Nenadic. 2015. Combining knowledge- and data-driven methods for de-identification of clinical narratives. *Journal of Biomedical Informatics*, 58:S53–S59. Supplement: Proceedings of the 2014 i2b2/UTHealth

Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

M. Douglass, Gari D. Clifford, A. Reisner, George B. Moody, and Mark Rg. 2004. Computer-assisted de-identification of free text in the mimic ii database. *Computers in Cardiology, 2004*, pages 341–344.

Daniele Faraglia. Faker. `https://github.com/joke2k/faker`.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5829–5839, Florence, Italy. Association for Computational Linguistics.

James Gardner and Li Xiong. 2008. Hide: An integrated system for health information de-identification. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 254–259.

GDPR. 2018. General data protection regulation. Official website of the European Union.

Dilipkumar Gupta, Melissa I. Saul, and John R. Gilbertson. 2004. Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. *American journal of clinical pathology*, 121(2):176–86.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234 – 1240.

Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu. 2015. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of Biomedical Informatics*, 58:S47–S52. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42. Supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

Omri Mendels and Avishay Balter. Presidio: Context aware, pluggable and customizable data protection and de-identification sdk for text and images.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Maxim Samsonov, Jim Geovedi, Jim Regan, György Orosz, Paul O'Leary McCann, Søren Lind Kristiansen, Duygu Altinok, Roman, Leander Fiedler, Grégory Howard, Wannaphong Phatthiyaphaibun, Explosion Bot, Sam Bozek, Mark Amery, Yohei Tamura, Björn Böing, Pradeep Kumar Tippa, Leif Uwe Vogelsang, Ramanan Balakrishnan, Vadim Mazaev, GregDubbin, jeannefukumaru, Jens Dahl Møllerhøj, and Avadh Patel. 2020. explosion/spaCy: v3.0.0rc: Transformer-based pipelines, new training system, project templates, custom models, improved component API, type hints & lots more.

Maximilian Mozes and Bennett Kleinberg. 2021. No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization.

Mitchell Marcus Eduard Hovy Sameer Pradhan Lance Ramshaw Nianwen Xue Ann Taylor Jeff Kaufman Michelle Franchini Mohammed El-Bachouti Robert Belvin Ann Houston Ralph Weischedel, Martha Palmer. 2013. Ontonotes release 5. Philadelphia: Linguistic Data Consortium.

Patrick Ruch, Robert H. Baud, Anne-Marie Rassinoux, Pierrette Bouillon, and Gilbert Robert. 2000. Medical document anonymization with a semantic lexicon. *Proceedings. AMIA Symposium*, pages 729–33.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*, 75:S4–S18. Supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 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. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. *Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium*, pages 333–7.

Sean M. Thomas, Burke W. Mamlin, Gunther Schadow, and Clement J. McDonald. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. *Proceedings. AMIA Symposium*, pages 777–81.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Ben Wellner, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leonid Peshkin, Alex Yeh, Janet Hitzeman, and Lynette Hirschman. 2007. Rapidly Retargetable Approaches to De-identification in Medical Records. *Journal of the American Medical Informatics Association*, 14(5):564–573.

Hui Yang and Jonathan M. Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30–S38. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC medical informatics and decision making*, 19(5):232.