When Is a Name Sensitive?
Eponyms in Clinical Text and Implications for De-Identification

Thomas Vakili, Tyr Hullmann, Aron Henriksson and Hercules Dalianis
Department of Computer and Systems Sciences
Stockholm University, Kista, Sweden
{thomas.vakili, aronhen, hercules}@dsv.su.se
tyrhullmann@gmail.com

Abstract

Clinical data, in the form of electronic health records, are rich resources that can be tapped using natural language processing. At the same time, they contain very sensitive information that must be protected. One strategy is to remove or obscure data using automatic de-identification. However, the detection of sensitive data can yield false positives. This is especially true for tokens that are similar in form to sensitive entities, such as eponyms. These names tend to refer to medical procedures or diagnoses rather than specific persons. Previous research has shown that automatic de-identification systems often misclassify eponyms as names, leading to a loss of valuable medical information. In this study, we estimate the prevalence of eponyms in a real Swedish clinical corpus. Furthermore, we demonstrate that modern transformer-based de-identification systems are more accurate in distinguishing between names and eponyms than previous approaches.

1 Introduction

De-identification of data invariably reduces information content by either removing, concealing, or replacing sensitive text with pseudonyms. Pseudonymization of data based on automatic identification and replacement of personally identifiable information (PII) may also introduce misleading information if tokens or text spans are erroneously misclassified as PII. Tokens are more likely to be misclassified as PII if they share common features with PII of a certain class. Such situations often arise in clinical texts, which often contain *eponyms* (Kucharz, 2020). These medical terms are named after a researcher or clinician, typically somebody involved in the discovery or invention or discovery of the phenomenon bearing their name. Sometimes, it can also be the name of a patient affected by a disorder. Since eponyms refer to medical phenomena rather than persons, they should not be considered sensitive.

It is believed that there are over 8,000 medical eponyms. As discussed by Kucharz (2020), eponyms can refer not only to diseases but to a wide range of categories including tests, surgical procedures and anatomical structures. These eponyms can cause difficulty when trying to automatically detect PII. In one study, it was shown that while only 0.81% of clinical entities were misclassified as PII, this was substantially higher for eponyms, where between 10 and 49% of eponyms were misclassified as PII (Meystre et al., 2014).

The following example highlights the problem: *Dr. Sjögren suspects the patient has Sjögren's syndrome*. In this example, *Sjögren's syndrome* is an eponymous disorder which is being treated by a physician who happens to have the same name. When de-identifying the sentence, *Sjögren* should be concealed in *Dr. Sjögren* but not in *Sjögren's syndrome*. Concealing the eponymous name of the syndrome removes clinical information which could potentially be very important for the intended users of the data. However, it is not clear how prevalent eponyms are in clinical text and to what extent transformer-based named entity recognition (NER) systems trained to identify PII can distinguish between eponyms and sensitive names.

In this study, we estimate the prevalence of eponyms in a large corpus of Swedish clinical text. We also create a manually annotated corpus of clinical notes containing one or more eponyms and use this corpus to study the extent to which classifications of names overlap with eponyms. To that end, we employ a NER system trained to detect sensitive entities (e.g., names). In other words, we seek to understand how eponyms affect these models' ability to distinguish between actual names and eponyms. The main contributions of this study are summarized below:

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 76–80
March 21, 2024 ©2024 Association for Computational Linguistics

- We estimate that around 0.04% of tokens in clinical notes are eponyms and that these have a slight tendency to cluster in the same notes.

- We show that modern NER systems based on BERT are less likely to misclassify eponyms than older systems evaluated in previous studies.

- We discuss the implications of eponyms for automatic de-identification of clinical text and data utility.

- We create a clinical corpus annotated with eponyms that we plan to de-identify and make available to researchers.

2 Related Research

Research looking specifically at eponyms is scarce. The studies that are available often focus on the intersection of the de-identification of clinical texts and the detection of disorders. Berg et al. (2020) performed de-identification experiments and observed that rare eponyms in the training data tended to be misclassified as last or first names to a very high degree, but there were also cases where eponyms in the training data were misclassified as last or first names. Meystre et al. (2014) compare five de-identification systems and their flaws in erroneously detecting eponyms as protected health information (PHI)[1] in American clinical text. Three systems (MIT, MIST and HIDE) misclassify approximately 10% of all eponyms as PHI, and the other two systems (HMS and MEDs) misclassify as many as 40% of all eponyms as PHI.

Berg et al. (2020) created an eponym lexicon by using a NER model for clinical entities, i.e., a system to identify *Findings*, *Disorder*, *Body Parts* and *Drugs* in a Swedish clinical text. Then, they investigated whether these were based on the name of a person, in which case it was marked up as an eponym and added to the eponym lexicon. Finally, the created lexicon was manually reviewed to ensure correctness. The resulting eponym lexicon contains 275 eponyms.

Several studies have examined the impact of de-identification on data utility for machine learning. Results are highly contingent on an appropriate sanitization algorithm and a sufficiently strong NER model for detecting sensitive data (Berg et al., 2020; Lothritz et al., 2023). However, there are several examples of studies showing that data utility can be maintained for both fine-tuning, pretraining, and combined scenarios (Vakili and Dalianis, 2022; Verkijk and Vossen, 2022; Vakili et al., 2023). These studies examine the impact of de-identification by evaluating models trained to perform downstream tasks. A shared limitation is that these studies study the impact on downstream task performance overall. As such, these studies cannot conclusively rule out that there may be other scenarios where de-identification could still have a disparate impact on data utility. For instance, misclassifying and removing information contained in eponyms could be harmful in many scenarios, and examining this specific risk provides deeper insights into possible pitfalls in de-identifying clinical data.

3 Data and Experiments

In this study, we estimate the prevalence of eponyms in a large sample of Swedish clinical texts by using an eponym lexicon to automatically identify mentions of eponyms in clinical notes. We use this to create an eponym corpus by randomly sampling 1,000 clinical notes with at least one detected eponym mention. These notes are then manually reviewed and corrected while we also calculate inter-annotator agreement among four annotators. Finally, we fine-tune a Swedish clinical BERT model to identify PII and calculate to what extent eponyms are misclassified.

3.1 Creating an Eponym Corpus

The data used in this study was the Stockholm Eponym Corpus[2]. This is a subset extracted from the research infrastructure Health Bank[3] (Dalianis et al., 2015), which contains over 2 million patient records from the years 2007-2014 from over 500 clinical units. The data originates from the Karolinska University Hospital in Stockholm, Sweden. The eponym lexicon created by Berg et al. (2020) was used to find eponyms in the clinical text.

The total number of detected tokens and eponyms can be seen in Table 1. In the corpora, approximately 0.04% of all tokens are eponyms. For scale, this can be compared with the prevalence of

[1] PHI are a form of PII specified by the American HIPAA regulation.

[2] This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

[3] Health Bank, https://www.dsv.su.se/healthbank

Corpora	Tokens	Flagged eponyms	Estimated eponyms
Health Bank Subset (1%)	27,837,617	12,066	11,016
Entire Health Bank	∼2,800,000,000	N/A	∼1,108,000

Table 1: The number of flagged eponyms (based on the matching algorithm) and the estimated minimum number of real eponyms (based on the precision of the algorithm).

Term	Occurrences
Babinski(s)	1869
Romberg(s)	1777
Grasset(s)	1325
Crohn(s,'s)	944
Parkinson(s,'s)	738
Alzheimer(s,'s)	490
Sjögren(s)	475
Donder(s)	351
Valsalva	322
Lasegue(s,é)	295
Graves('s)	290
Akilles	256
Raynaud(s,'s)	217
Bechterew(s)	216
Whipple(s)	173
Willebrand(s)	169
Wegener(s)	162
Waldenström(s)	176
Robin(s)	179
Dix	154

Table 2: Top 20 highest occurrences of the eponyms from the Stockholm Eponym Corpus, including spelling variants.

PII which has been estimated as being two to four times more common (Dalianis, 2018).

Due to computational constraints, one percent of the Health Bank corpus was randomly extracted for the experiments. This subcorpus consists of 1,402,782 notes containing 27,837,617 tokens and was tagged for eponyms using exact matching with the eponym lexicon. In total, 9,795 notes were flagged as containing eponyms, and 12,066 matching eponyms were found, as shown in Table 1.

Out of the 9,795 notes with eponyms, 1,000 notes containing the eponym tag were randomly extracted for manual annotation. The order of the notes was randomized before being split into five subsets of 200 notes. These notes were manually annotated by four annotators. Each annotator was assigned 400 notes, 200 of which were unique and 200 that were shared. The resulting 1,000 notes corpora is called the Stockholm Eponym Corpus. The inter-annotator agreement (IAA) was determined using the Krippendorff's alpha (Krippendorff, 1970) and was calculated as 0.97 for the 200 samples annotated by all four annotators.

These 200 shared samples were then used to estimate the precision of the eponym lexicon. After resolving the disagreements between the annotators, the precision was determined as 0.913. No attempts were made to estimate the recall of the matching algorithm, as the annotated samples were only selected from the subset in which the algorithm had found eponyms. Based on the precision of the matching algorithm, a lower bound for the total number of eponyms in the Health Bank was estimated and listed in Table 1.

During the manual annotation, new eponyms were discovered, annotated, and added to the lexicon. This process led to extending the eponym lexicon from 275 eponyms to 317 eponyms. The updated eponym lexicon was used for the final matching presented in Table 1 and 2, respectively.

3.2 Evaluating Misclassification of Eponyms

Previous studies have shown that NER systems for classifying PII tend to have lower precision for tokens that are eponyms. To study this, a BERT-based NER model was trained using the Stockholm EPR PHI Corpus (Dalianis and Velupillai, 2010). This corpus covers a range of PII classes and consists of 380,000 tokens, of which 4,800 are PII. Crucially, it covers both first and last names – entity types that are commonly associated with eponyms. A Swedish clinical BERT model called SweDeClin-BERT (Vakili et al., 2022) was used as the base model. The fine-tuned NER model was then used to tag the corpus described in Section 3.1, creating a version containing both tags for PII and eponyms.

The new version of the corpus, which contained parallel tags for eponyms and PII, was examined to determine how often eponyms were misclassified as PII. A total of 82 tokens out of the 1,319 tokens annotated as eponyms were classified as PII. In other words, approximately 6.2% of eponyms were misclassified. Interestingly, the NER tagger did not only confuse eponyms with names but also with locations and organizations. Statistics for the misclassifications are shown in Table 3.

PII tag	Misclassified Eponyms	Non-Eponym Classifications
Last Name	72	227
First Name	7	254
Organization	2	14
Location	1	58

Table 3: Many PII were predicted in the Eponym Corpus. Some of these were eponyms. Eponyms were misclassified mainly as names and, in a few cases, as locations or organizations.

4 Discussion and Conclusions

4.1 Observations During Annotation

One observation during the annotation process was that eponyms were rarely present in the same context or sentence as PHI. In other words, the scenario showcased in the example in the introduction was uncommon. Eponyms often occur in bursts in the text, in discussions of possible disorders, or in descriptions of tests that had been conducted. This phenomenon is illustrated in Figure 1.

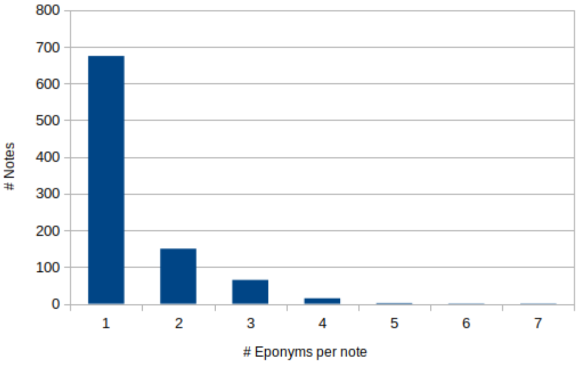

Figure 1: Although the majority of notes contain just one eponym, nearly half of all detected eponyms occurred in notes containing at least one additional eponym.

There were some examples where either the clinician's name or the patient's name coincided with the eponym. *Robin* was present in the eponym lexicon to catch references to Robin's syndrome, but these mentions were more often misclassified as eponyms since Robin is a common Swedish name.

Many names are also non-eponymous words. For example, *Still* was in the eponym lexicon but was also a common non-eponymous word (with the same meaning as in English e.g., *to sit still*).

4.2 Improvements Over Previous Research

The results highlighted in Section 3.2 indicate that the problem of eponyms being misclassified as PII is less prevalent in our study compared to previous research. In particular, the outcome can be contrasted to the results of Berg et al. (2020), who also used data from the Health Bank. It is difficult to confidently conclude what these differences are caused by. One hypothesis is that transformer-based models better capture the context surrounding a token. This could allow them to better distinguish when a name is used as a name and when it is used as an eponym. Indeed, these uses are grammatically distinct and are often obvious to a human observer. Further experiments would be needed to conclusively ascribe the differences in results to this capability or determine if they are due to other factors.

4.3 Conclusions

Protecting privacy is crucial in the clinical domain but also comes with domain-specific challenges. Eponyms contain valuable clinical information and we estimate, based on our results, that at least 0.04% of all tokens in clinical notes are eponyms. Previous research has found that automatic de-identification systems can struggle to distinguish between eponyms and actual private names that need to be sanitized. Our results show that modern transformer-based NER models, such as those based on BERT, are more effective in separating these two forms of names. This study also presents a new annotated corpus containing a wide range of eponyms. We plan to release a de-identified version of this resource once the necessary ethical permissions have been obtained.

5 Limitations

While three of the four annotators had prior experience working with clinical text, none were trained medical professionals but computer scientists. Some eponyms may have been missed during the annotation process, and others may have been erroneously annotated. In cases where the annotators needed clarification, they searched online for sources indicating whether or not a name was an eponym. The high IAA indicates that the annotations are reliable, but the lack of medical expertise limits the extent to which the annotations can be trusted.

A related issue is that the eponym corpus is not a random sample of the entire Health Bank. Instead, it is a consciously chosen subset that was deemed highly likely to contain eponyms based on the matching algorithm described in Section 3.1. Starting from a purely random subset of the Health Bank could have led to more robust results and would have allowed us to calculate the recall for the matching algorithm. This was not deemed feasible due to the very low prevalence of eponyms in the overall corpus. Starting from a random sample would have required far more annotators than were available for this project.

The risk of misclassifying eponyms was only examined for the SweDeClin-BERT model. It is possible that other architectures and models trained on other datasets may perform better or worse. Further research could benefit from including a more diverse range of models, including generative models. Nevertheless, the results of this study show that transformer-based models can be less affected by the misclassification risks than models described in earlier studies. Determining the mechanism behind this greater resilience is an interesting topic for future research.

Acknowledgement

We want to thank Hanna Berg for providing details about the eponym lexicon. We are also grateful for the helpful comments and suggestions from the reviewers. Finally, we want to thank the DataLEASH project and Digital Futures for funding this research.

References

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK- A Workbench for Data Science Applications in Healthcare. In *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015)*, pages 34–44. CEUR Workshop Proceedings.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, 1(1):6.

Klaus Krippendorff. 1970. Bivariate Agreement Coefficients for Reliability of Data. *Sociological Methodology*, 2:139–150. Publisher: [American Sociological Association, Wiley, Sage Publications, Inc.].

Eugeniusz Józï Kucharz. 2020. Medical eponyms from linguistic and historical points of view. *Reumatologia*, 58(4):258–260.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the Impact of Text De-Identification on Downstream NLP Tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands. University of Tartu Library.

Stéphane M Meystre, Oscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014. Text de-identification for privacy protection: a study of its impact on clinical text information content. *Journal of biomedical informatics*, 50:142–150.

Thomas Vakili and Hercules Dalianis. 2022. Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 383–388, Dublin, Ireland. Association for Computational Linguistics.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2023. End-to-End Pseudonymization of Fine-Tuned Clinical BERT Models.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252.

Stella Verkijk and Piek Vossen. 2022. Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1098–1103, Marseille, France. European Language Resources Association.