The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text

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

Abstract

The impact of de-identification on data quality and, in particular, utility for developing models for downstream tasks has been more thoroughly studied for structured data than for unstructured text. While previous studies indicate that text de-identification has a limited impact on models for downstream tasks, it remains unclear what the impact is with various levels and forms of de-identification, in particular concerning the trade-off between precision and recall. In this paper, the impact of de-identification is studied on downstream named entity recognition in Swedish clinical text. The results indicate that de-identification models with moderate to high precision lead to similar downstream performance, while low precision has a substantial negative impact. Furthermore, different strategies for concealing sensitive information affect performance to different degrees, ranging from pseudonymisation having a low impact to the removal of entire sentences with sensitive information having a high impact. This study indicates that it is possible to increase the recall of models for identifying sensitive information without negatively affecting the use of de-identified text data for training models for clinical named entity recognition; however, there is ultimately a trade-off between the level of de-identification and the subsequent utility of the data.

1 Introduction

There is a growing demand for access to large amounts of healthcare data in order to facilitate research and development of tools for healthcare management and clinical decision support, not least as a result of the increasing application of AI and machine learning in healthcare. However, to enable large-scale secondary use of sensitive healthcare data, there is a need for automatic privacy-protecting methods for clinical text; manual de-identification to ensure that the data does not contain personal information is often prohibitively expensive.

Privacy-protecting methods for de-identification of data generally address three privacy risks: (i) the risk that someone's records can be uniquely identified in a dataset, (ii) preventing linkage from one dataset to another, and (iii) the risk of inferring sensitive information about an individual from the dataset (EU, 2014). De-identification of unstructured clinical text commonly focuses on identifying potentially identifiable information within predefined classes in an approach based on named entity recognition (NER), where protected health information (PHI) is first identified and subsequently obscured in some fashion (Meystre et al., 2010). This technique addresses the first of the mentioned privacy risks, and is the focus of this study.

De-identification techniques for structured data generally address all three mentioned privacy risks, and their impact on data quality and data utility are more thoroughly studied than the impact of de-identification on unstructured text (Iwuchukwu et al., 2007). Research on structured data shows that de-identification techniques applied to structured data may lead to reduced data quality (Xia et al., 2015). For both structured and unstructured data, information necessary for answering a research question may be identifying, and therefore required to be removed or altered to ensure the privacy of data subjects. Furthermore, for automatic de-identification, relevant information may be changed or removed unintentionally due to non-sensitive tokens being misclassified as sensitive. Studies have indicated that de-identification may erroneously alter or hide data significant for other tasks; however, no significant impact has so far been observed when comparing models for downstream tasks trained with original versus de-identified text data (Deleger et al., 2013; Obeid et al., 2019; Meystre et al., 2014). It is, however,

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 1–11

https://doi.org/10.18653/v1/P17

still not clear how the impact is related to various forms and levels of de-identification, in particular the trade-off between the precision and recall of the de-identification system.

Common measures for evaluating de-identification models for unstructured data are precision, recall and F_1-score. High recall is generally preferred over high precision, as the privacy of the data subjects is prioritised over potential loss of document interpretability (Ferrández et al., 2012). There is, at the same time, a concern that poor precision would have a negative impact on the quality of the data. In practice, precision and recall are typically evaluated as equally important when using F_1-score as the primary evaluation metric. However, if precision of the de-identification model is of less importance and it turns out that this does not have a clear impact on using the data for some downstream task, it would entail that de-identification systems can be adapted for high coverage – but also that recall should carry more weight than precision when evaluating de-identification systems.

On the other hand, if de-identification impacts the text data quality negatively, the concern is that it would lower the possibility to use the data for building models for various downstream tasks, i.e. data utility would decrease once it has been de-identified. Examples of downstream tasks are detection of healthcare-associated infections, adverse drug events and early cancer symptoms.

This study intends to answer the question of how the precision of a de-identification system affects data quality and, in particular, data utility by investigating how different levels of de-identification affects the performance of downstream clinical NER. In particular, the study aims to investigate the impact of the trade-off between precision and recall, as well as different methods for concealing PHIs.

2 Related Research

To our knowledge, no studies have specifically focused on the trade-off between precision and recall and its impact on downstream tasks. There are, however, studies that have investigated the impact of de-identification.

A study by Meystre et al. (2014) showed that de-identification reduces the information content, and leads to the possible introduction of misleading information if tokens are replaced with pseudonyms. In the study, only 0.81% of clinical named entity annotations were erroneously detected as PHI, but between 10 and 49% of all eponyms[1] were misclassified as PHI. Fewer SNOMED-CT concepts were also found, but this was largely explained by incorrectly labelled SNOMED-CT concepts in the original dataset. Studies on downstream tasks have, however, not shown that de-identification has a negative impact. No significant differences could be seen between using original text or de-identified text as training data for clinical text classification (Obeid et al., 2019). In contrast, one study observed potentially significant benefits of training on de-identified data for medication name extraction (Deleger et al., 2013). The reduction of dimension in de-identified text has been hypothesised to potentially improve machine learning performance (Obeid et al., 2019).

While previous studies have compared different systems with various levels of precision (Obeid et al., 2019; Meystre et al., 2014), no study has explicitly compared the impact the precision of a de-identification system has on downstream tasks. In Deleger et al. (2013), a recall bias was introduced in two versions and those models were compared to the original version. The systems with a slightly lower precision performed similarly to the original one in terms of performance on a medication name extraction task.

3 Methods and Materials

In this paper, the impact of various forms of de-identification – where a trade-off is made between precision and recall of the model used for identifying PHI, as well as using different levels of de-identification – on downstream NER tasks, in this case identifying various clinical entity types, is studied.

This paper uses four corpora: one for development of de-identification models and three for evaluating the impact of de-identification on downstream clinical NER. The method is presented first, followed by a description of the corpora. The experiments include five models for identifying PHI, each one trading off precision for increased recall to different extents, as well a four concealment strategies that hide the identified PHI to different degrees. These PHI models and concealment strategies are used to de-identify three clinical corpora[2],

[1] Eponyms are terms named after researchers, such as *Crohn's disease*, *Cushing's syndrome* or *Waldenström macroglobulinemia*.

[2] This research has been approved by the Swedish Ethical

which, in turn, are used for training downstream clinical NER models. The impact of using different PHI models and concealment strategies is then analysed from a number of different perspectives.

3.1 Methods

The impact of PHI models and concealment of PHI is evaluated based on their performance as training data for downstream clinical NER tasks. The overlap and co-occurrence between manually annotated clinical entities and predicted PHI are also analysed to find out if certain PHI classes have a bigger impact on the downstream clinical NER tasks. Furthermore, the impact of de-identification on eponyms are analysed on a new corpus specifically made for this purpose.

3.1.1 De-identification Process

The de-identification process consists of two main steps: (i) identification of PHI and (ii) concealment of PHI.

3.1.2 Optimising the F-score

Optimisation for different F-scores was carried out to obtain models with higher recall at the expense of precision, with the aim of investigating the effect this has on downstream clinical NER. Deleger et al. (2013) introduced a recall bias by changing the predicted label of non-PHI tokens with system-generated probability less or equal to a threshold of 0.95 to the PHI label with the second highest probability. In this study, this was done in a similar fashion by changing predicted non-PHI labels to their most likely alternative PHI label if the the marginal score was lower than the set threshold. The marginal probability specifies the model's confidence in predicting each label of an input sequence, without regard to the outcome of other variables (Sutton and McCallum, 2012).

The thresholds were decided through grid search, optimising toward five different F-scores: F_1, F_4, F_{10}, F_{20} and F_{40}. F-score is the weighted mean of precision and recall, see Eq. 1. For F_1, equal weight is given to precision and recall. For F_2, recall is given twice the weight of precision. Hence, β will obtain the following values: 1, 4, 10, 20 and 50.

$$F\text{-}score : F_\beta = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \quad (1)$$

Review Authority under permission no. 2019-05679.

3.1.3 Concealment Strategies

The next step is to conceal the PHI. The four concealment strategies used in this study are: *Pseudo*, *Class*, *Mask* and *Remove*.

- *Original* – No de-identification method. Example: "Eva slept."
- *Pseudo* – To replace the identified PHI with a surrogate. Example: "Mary slept."
- *Class* – To replace the identified PHI with the PHI class. Example: "<First_Name> slept."
- *Mask* – To replace the identified PHI with *XXXX*. Example: "XXXX slept."
- *Remove* – To completely remove the identified PHI and the sentence it is contained. Example: " "

The different methods hide information to varying degrees. *Pseudo* replaces some information, but, for example, may retain information about time between different events by shifting dates consistently. For *Class*, there is still information about the type of PHI found, which is missing for *Mask*. *Remove* removes not only the PHI itself but also the context surrounding it, i.e. all sentences where a PHI is found are removed.

The pseudonymisation algorithm was similar to the one described in (Dalianis, 2019; Berg et al., 2019). The changes are based on the error analysis in (Berg et al., 2019). The changes are:

- Uncommon names are replaced with uncommon names.
- The number of tokens of each PHI instance is kept.
- The format of dates are kept.
- Tokens for *Health Care Units* are replaced with the same strategy as the one used for replacing. *Locations* in Dalianis (2019).
- Not all *Health Care Units* tokens are replaced, but at least one token within every entity. Locations and named entities are replaced, while for example tokens like "hospital" or "clinic" are not replaced,

For each clinical NER corpus, 25 dataset variants were produced. The datasets had six levels of de-identification based on f-scores, ranging from not de-identified to increasingly larger percentages of both true and false positives. The identified PHI were then managed using four different concealment strategies: pseudonymising the PHI (*Pseudo*), replacing the PHI with the classified PHI

type (*Class*), masking the PHI (*Mask*) or removing whole sentences containing PHI (*Remove*). This resulted in 24 de-identified datasets and one original dataset for training.

3.2 Materials

All corpora are contained in the research infrastructure Health Bank – the Swedish Health Record Research Bank[3]. Health Bank contains electronic patient records from over two million patients from Karolinska University Hospital from the years 2006–2014.

PHI Corpus

The de-identification models was trained, as well as evaluated, on the Stockholm EPR PHI Corpus (Velupillai et al., 2009; Dalianis and Velupillai, 2010). The Stockholm EPR PHI Corpus consists of 98 patient records in Swedish from five clinical units at Karolinska University Hospital: *neurology, orthopaedia, infection, dental surgery* and *nutrition*. The corpus contains nearly 200,000 tokens in total. The annotated classes are: *Age, Full Date, Date Part, First Name, Last Name, Health Care Unit, Location* and *Phone Number* distributed over 4,826 annotated entities. The dataset includes free text and information about which section the text is from. The distribution is imbalanced with roughly 3% of all tokens being part of an annotated entity.

Four out of five clinical units were used for training and the tuning of the hyper parameters. The data from the *neurology clinical unit* was used for the final evaluation.

Clinical Entity Recognition Corpora

The following three corpora were used for building and evaluating clinical NER models, with and without de-identification.

Stockholm EPR Clinical Entity Corpus is a corpus in Swedish for clinical NER with clinical notes from an internal medicine emergency unit at Karolinska University Hospital (Skeppstedt et al., 2014). The corpus was annotated by three annotators; the annotation process and the resulting corpus are described in (Skeppstedt et al., 2014; Kvist et al., 2011). The dataset is annotated with the labels: *Explicit Disorder*, *Implicit Disorder*, *Finding*, *Drug* and *Body structure* distributed over totally 7,946 annotated entities.

Stockholm EPR (Adverse Drug Event) ADE Corpus is a corpus in Swedish for adverse drug events, annotated with clinical named entities (Henriksson et al., 2015). The clinical notes are extracted from the larger Health Bank by extracting data with ICD-10 codes marking an adverse drug event. The dataset uses the labels *Disorder*, *Finding*, *Drug*, *ADE Cue* and *Body structure* distributed over totally 3,789 annotated entities.

Stockholm EPR Cervical Cancer Corpus is a corpus in Swedish with clinical records for patients with a cervical cancer diagnosis. The texts are annotated with the labels *Finding*, *Disorder* and *Body part*, distributed over totally 7,663 annotated entities. The annotation process is described in (Weegar et al., 2015).

Eponym Corpus

An additional corpus was constructed in order to enable a more thorough analysis of the impact on eponyms. The dataset is a subset of a larger corpus extracted from Health Bank, with over 213 million tokens. For each of the 12 most common eponyms, one hundred sentences in which the eponym appeared were extracted.

3.3 Experimental setup

In the analyses of the clinical NER corpora, *First Name* and *Last Name* are merged to *Person* and *Date Part* and *Full Date* to *Date*. *Explicit disorder*, *Implicit Disorder* and *Disorder* are similarly all merged to one category. In the analyses, we present results averaged over the three corpora.

The PHI models are evaluated on the subset of Stockholm EPR PHI Corpus that they are not trained on. The main evaluation is a binary evaluation to investigate how many non-PHI that are classified as PHI. The evaluation is also token-based, meaning that they are evaluated on a per-token basis. The system's ability to locate where an PHI begins and another ends is not of great importance for the replacement step, but instead the focus is on whether tokens are replaced or not.

For each clinical NER corpus, the system trained for finding clinical entities was basic. CRFSuite was used with word features and orthographic features. Due to the large number of datasets no hyperparameter tuning was done. The hyperparameter's used are the default parameters for CRFSuite, but with an added *c1* of *0.1* and a *c1* of *0.2*. 5-fold cross validation was used.

The overlap of classified clinical entities and PHI are investigated, to see how they relate to each other. For *class*, *mask* and *pseudo* this overlap leads to a

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

loss of information and training examples for the clinical NER task. The overlap is analysed on a token level, since one PHI entity may span multiple clinical entities, or only parts of a clinical entity.

In comparison to the other concealment strategies, the de-identification's precision with *Remove* has a clear impact in terms of the overlap between clinical entities and PHI. Sentences with a co-occurrence of at least one token classified as a PHI and an annotated clinical entity are removed with this method. This means that there are fewer examples of clinical entities in the datasets de-identified with *Remove* as the concealment strategy. In the co-occurrence analysis, the co-occurence is measured to get information about how the identification of PHI affects the clinical entity information when using the *Remove* method.

4 Results

The results for PHI identification cross-validated on the PHI corpus are first presented to provide an estimate of the performance of each PHI model. The impact of de-identification on the three clinical NER corpora used for the downstream tasks are then analysed in three different ways: (i) the overlap between predicted PHI and manually annotated clinical entities, (ii) the co-occurrence of predicted PHI and manually annotated clinical entities in the same sentences, and (iii) the impact on downstream clinical NER performance. Finally, the misclassification of eponyms as PHI is specifically studied.

4.1 PHI Identification

The results from the development of PHI models optimised for different F-scores are presented below. The best combination of thresholds for each F-score, F_1, F_4, F_{10}, F_{20} and F_{40}, are presented in Tables 3 and 4 in the Appendix.

The cross-validated performance scores of the PHI identification models with adjusted marginal scores are shown in Table 1. As expected, with increased bias, recall increases at the expense of precision. While the recall improves from *model* F_1 to *model* F_{40} by a total of 7 percentage points, the precision drops by as much as 83 percentage points. As expected, the F_1-score is highest when using the model optimised for F_1. The highest number of false positives was observed for the class *Health Care Unit*.

PHI Model	P	R	F_1
F_1	96.07	92.82	94.41
F_4	77.82	97.55	86.57
F_{10}	44.95	99.53	61.93
F_{20}	26.47	99.72	41.83
F_{40}	12.74	99.94	22.60

Table 1: Token-based binary evaluation of the de-identification (PHI) models with differently set marginal scores, optimised for a particular F-score. P stands for precision, R for recall and F_1 for F_1-score.

4.2 Overlap Analysis

Only 1% of clinical entities are affected by the de-identification process for F_1, in terms of partial overlap with PHI entities. As expected, a larger recall bias leads to an increasing amount of tokens being classified as PHI. As many as around one third of all clinical entities are classified as PHI by the F_{40} model. As can be seen in Figure 1, *Health Care Unit*, in comparison to other PHI classes, overlaps more with clinical entities for the models with low precision, while *Person* overlaps more with other PHI in the models with high precision. In relation to the number of classified cases, *Location* is, however, the PHI class that overlaps the most with the clinical entities (26% of all classified locations).

The different clinical entities are affected to different degrees. As can be seen in Figure 1, the clinical entity that overlaps the most with PHI is *Drug*, and the one that overlaps the least with PHI is *Finding*.

4.3 Co-occurrence Analysis

As can be seen in Figure 2, the *Remove* concealment strategy leads to the removal of a large amount of clinical entities. As expected, the PHI identification models with lower precision and higher recall removes more PHI compared to the other models. For F_1, F_4, F_{10} and F_{20}, a greater degree of sentences without clinical entities are removed than with clinical entities. For F_{40}, there is an equal amount removed.

While there is little overlap between the PHI class Age and clinical entities, a disorder is mentioned in 67% of sentences with an age annotation in F_1, for example *82 year old man with Alzheimers*. With the *Remove* concealment strategy, these mentions of disorders will be removed from the de-identified dataset.

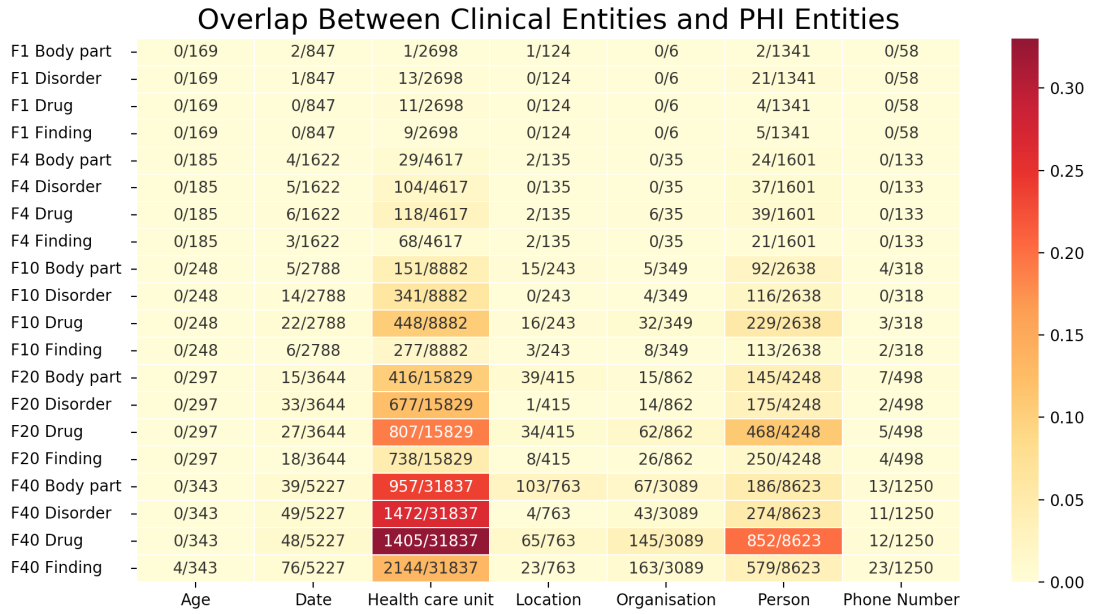

Figure 1: Presentation of clinical entities replaced during de-identification and which label they were classified as. The number shown for each index is the how many overlaps per identified PHI in total, and the colour of each column represents the percentage of clinical entities (see left column) that are overlapped for each clinical entity class.

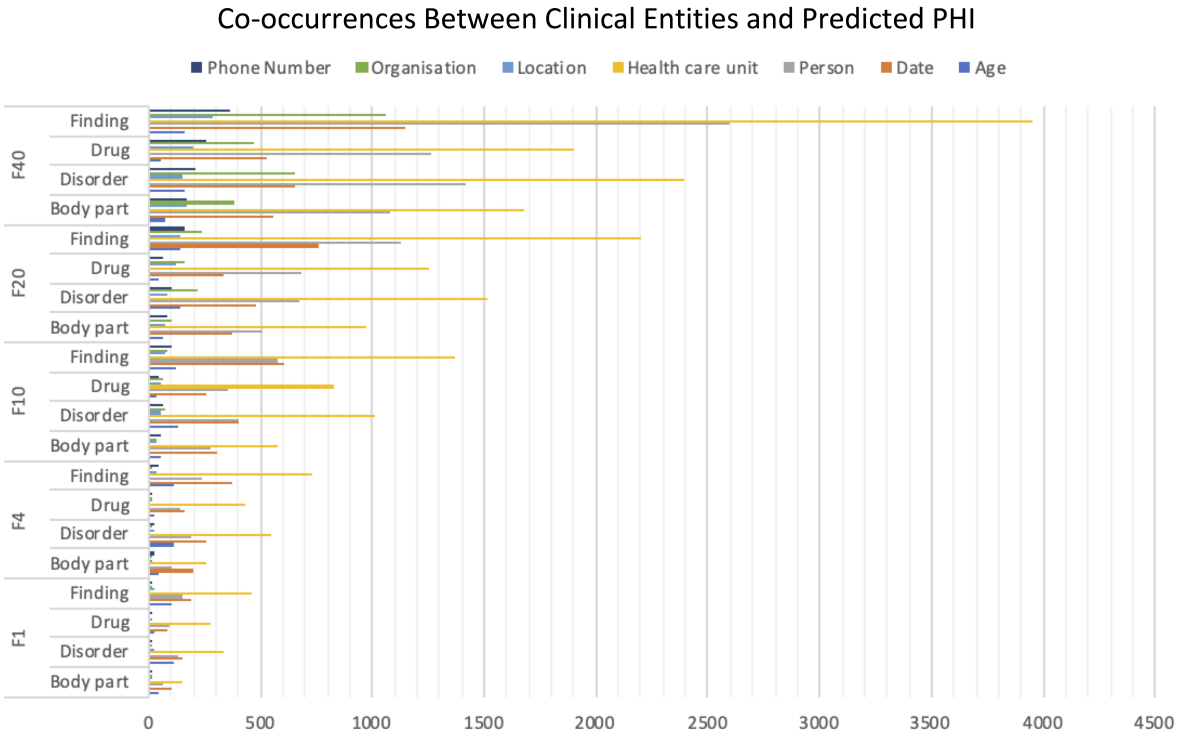

Figure 2: The figure presents the number of sentences with a co-occurrence of a clinical entity and a predicted PHI.

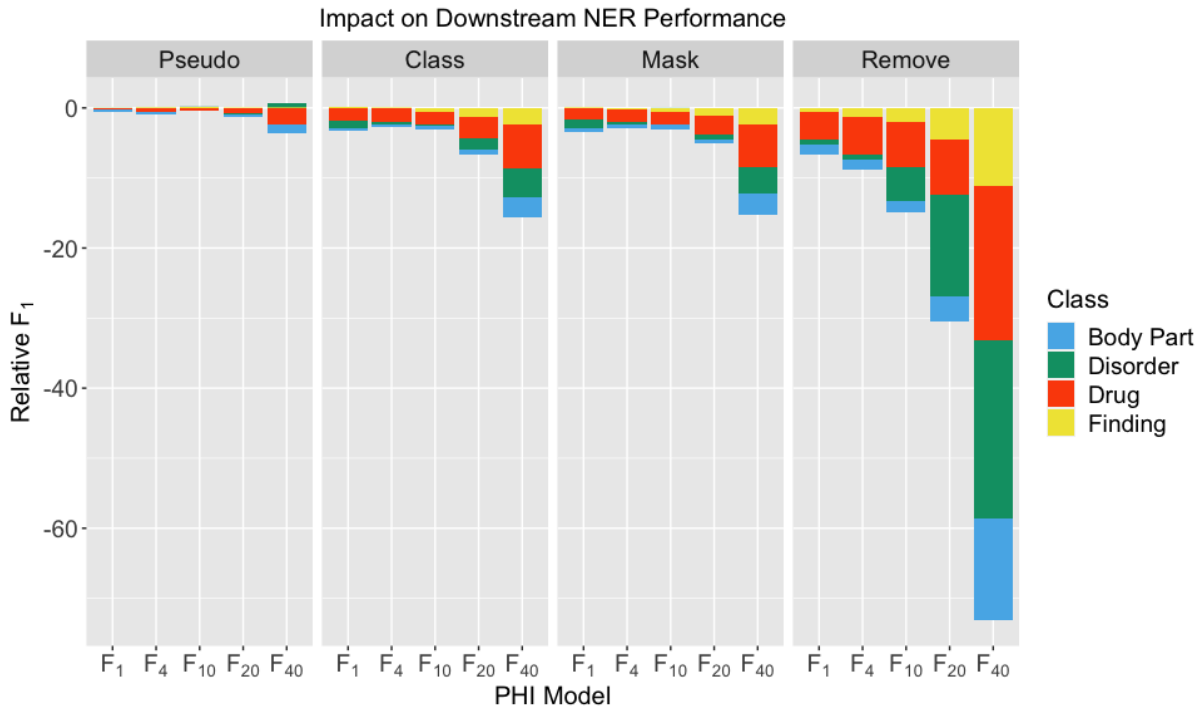

Figure 3: The impact on predictive performance, in terms of relative F_1-score, when training clinical NER models with de-identified data (compared to without de-identification). Results for the target clinical entity classes are shown for five different PHI identification models and four different concealment strategies.

4.4 Downstream NER Performance

As can be seen in Figure 3, the choice of concealment strategy for de-identification has a large impact on downstream clinical NER performance. With the *Pseudo* concealment strategy, the overall performance across models is fairly limited, whereas the impact is somewhat larger with the *Mask* and *Class* concealment strategies, with very little difference between them. A much bigger, negative impact is observed when applying the *Remove* concealment strategy.

The impact on the downstream clinical NER tasks are also different across PHI models. As expected, the performance tends to get increasingly worse the more the PHI model is trained to prioritise recall over precision. However, with *Pseudo*, *Class* and *Mask*, the differences are fairly small for the F_1, F_4 and F_{10} PHI models. With *Remove*, on the other hand, the differences across PHI models are markedly more pronounced.

The impact on performance is not equal across clinical NER classes. Overall – across PHI models and concealment strategies – the most negative impact was observed for the *Drug* class (-3.9%), followed by the *Disorder* class (-3.0%), the *Body Part* class (-1.7%) and the *Finding* class (-1.4%). Across concealment strategies, the *Drug* class was almost invariably the most impacted clinical entity, with the exception of the F_{20} PHI model, with relative F_1-scores ranging from -1.9% to -9.1%.

The least impacted class varied across PHI models, but was mostly *Body Part* or *Finding*, with relative F_1-scores ranging from -0.4% to -5.4%. In almost all cases, a monotonic decrease in performance is observed as the PHI models are giving increasing priority to recall at the expense of precision. Across PHI models, a similar pattern is observed, with the performance on the *Drug* class being negatively affected the most, ranging from -0.8% to -9.1%. However, with *Remove*, there is a greater impact on the *Disorder* class than the *Drug* class. For all concealment strategies, the *Finding* class is the least affected.

4.5 Eponym Analysis

According to previous studies, medical eponyms risk being mistaken as identifiable information, as they are derived from a personal name.

Throughout the three corpora, 21 different eponyms occurred, with a total of 57 mentions. With 43% of eponyms only being mentioned once, it would not be possible to make any conclusions based on that data. Therefore, an eponym

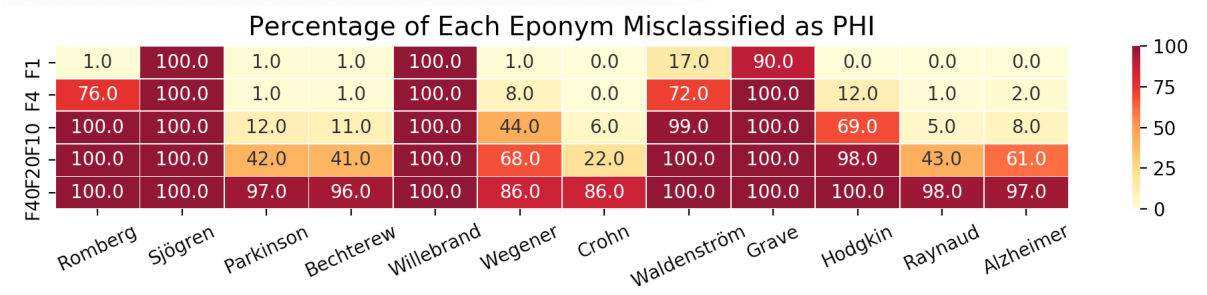

Figure 4: The figure shows how many percent of each eponym that are missclassified as PHI in the Eponym Corpus for different five PHI identification models, F_1, F_4, F_{10}, F_{20} and F_{40}.

dataset was created to be able to study the effect on eponyms further.

In the training data from the PHI corpus, *Parkinson* occurs 11 times, *Bechterew* 4 times, *Waldenström* 2 times, *Alzheimer* 1 time and *Romberg* 1 time. As shown in Figure 4, the eponyms that occurred in the training data are affected to a lesser extent than those that did not. *Sjögren syndrome* and *von Willebrand disease* are regardless of context classified as surnames by all models, together with *Grave's disease*. *Hodgkin lymphoma* and *Raynaud syndrome*, are never mentioned in the training set, but are classified as surnames less often than for example *Waldenström*. This is despite *Waldenström* being in the training data. With the F_{40} PHI identification model, almost all eponyms are classified as PHI, where 60% are misclassified as *Last Name* and 15% as *First Name*.

To summarise the analysis, the overlap for eponyms with PHI is greater than for disorders in general. The eponyms most likely to be affected by de-identification are those which bear a resemblance to common Swedish last names.

5 Discussion

In contrast to the findings by both Deleger et al. (2013) and Obeid et al. (2019), a small negative impact was indeed observed regardless of the precision and recall of the PHI model used for de-identification. However, the impact is very small unless the F_{40} PHI model or the *Remove* concealment strategy is used.

The worst performing model in (Stubbs et al., 2015) had a token-based binary precision of 76% (and a recall of 52%). The only models with an estimated precision above this in our study is F_1 and F_4. Despite the clear difference in precision between F_1 and F_{10}, F_{10} gives a high recall 99.5% for PHI identification, see Table 1, and the concealment strategies *Pseudo*, *Class* and *Mask* still allow for good downstream results. The trade-off between the level of de-identification and data utility may vary depending on the information sensitivity of the data, and assessments should be made on a case-by-case basis. According to our experiments it is to a certain extent, possible to raise recall at the expense of precision without affecting downstream performance. It is also important to balance this with the chosen concealment method, as they affect the utility to varying degrees. For example, F_{40} may be appropriate to combine with *Pseudo*, but not with *Remove*. Another possibility would be to use different concealment methods for different PHI, since some are not as sensitive as others, and create a recall bias for some labels but not for others.

PHI models with low precision combined with *Pseudo* as concealment strategy seem to have a much smaller impact on downstream tasks. A potential cause may be that not all *Health Care Unit* tokens were affected by the pseudonymisation.

In this study, the de-identification process is performed after manual annotation process of the clinical entity corpora. If the annotation process was performed after the de-identification, and there was more data available than could be annotated, the *Remove* method may have less impact.

Based on this study, certain PHI classes have a higher risk of overlapping with relevant clinical information than others. We have, however, not investigated how the precision and recall of specific PHI labels affect downstream tasks. The analysis of eponyms, where 75% are classified as either *First Name* or *Last Name* by the F_{40} PHI identification model, indicates that certain classes may have a greater impact on downstream tasks than others. The eponym analysis also indicates that, while the downstream impact is small on the clinical NER

corpora, other tasks may be affected to a greater extent. As eponyms risk being misclassified as PHI, there may be a need, if the diseases are relevant to the task, to deal with them specifically, for example by adding rules to avoid being unrecognised as disorders.

This study has focused on utility for clinical entity recognition. Future research may investigate how de-identification impacts other downstream tasks. It may also be of interest to have people read through the texts and see if they are readable and possible to use for research in more qualitative research on electronic health records.

6 Conclusion

This study demonstrates that corpora de-identified using PHI models with a moderate to high precision lead to similar performance when used for downstream clinical NER tasks. The impact is, however, affected by both the choice of concealment strategy and the trade-off between precision and recall, in particular when the precision is low.

Optimising the PHI identification model for F_4 gives a relatively lower precision of 77.82% and a higher recall of 97.55%. Compared to using a standard F_1-optimised model for de-identification, this results in higher privacy and a relatively small negative impact on downstream clinical named entity recognition.

This study indicates that it is possible to increase the recall of models for identifying sensitive information without negatively affecting the use of de-identified text data for training models for clinical named entity recognition

Furthermore the overlap analysis showed that with lower precision, there is an increase of overlap between clinical information and automatically labelled PHI. Some PHI labels overlap with clinical entities more than others. Different clinical entities are also more likely to be affected than others, like *Drug*. Eponyms may also risk being misclassified as last names, and studies interested in those may need to take extra precautions to handle those properly.

References

Hanna Berg, Taridzo Chomutare, and Hercules Dalianis. 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.

Hercules Dalianis. 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.

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*, 1:6.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 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.

EU. 2014. Article 29 data protection working party, opinion 05/2014 on anonymisation techniques. *EU, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf*.

Óscar Ferrández, Brett R South, Shuying Shen, F Jeff Friedlin, Matthew H Samore, and Stéphane M Meystre. 2012. Generalizability and comparison of automatic clinical text de-identification methods and resources. In *AMIA Annual Symposium Proceedings*, volume 2012, page 199. American Medical Informatics Association.

Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of biomedical informatics*, 57:333–349.

Tochukwu Iwuchukwu, David J DeWitt, AnHai Doan, and Jeffrey F Naughton. 2007. K-anonymization as spatial indexing: Toward scalable and incremental anonymization. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 1414–1416. IEEE.

Maria Kvist, Maria Skeppstedt, Sumithra Velupillai, and Hercules Dalianis. 2011. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems- Future vision, a physician's perspective. In *Proceedings of 9th Scandinavian Conference on Health Informatics, SHI 2011, Oslo, (Eds.) Fensli and Dale, Tapir Academic Press*, pages 31–35.

Stéphane M Meystre, Óscar 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.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic De-Identification of Textual Documents in the Electronic Health Record: A Review of Recent Research. *BMC medical research methodology*, 10(1):70.

Jihad S Obeid, Paul M Heider, Erin R Weeda, Andrew J Matuskowitz, Christine M Carr, Kevin Gagnon, Tami Crawford, and Stephane M Meystre. 2019. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health technology and Informatics*, 264:283.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158.

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.

Charles Sutton and Andrew McCallum. 2012. An introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Rebecka Weegar, Maria Kvist, Karin Sundström, Søren Brunak, and Hercules Dalianis. 2015. Finding cervical cancer symptoms in Swedish clinical text using a machine learning approach and NegEx. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1296. American Medical Informatics Association.

Weiyi Xia, Raymond Heatherly, Xiaofeng Ding, Jiuyong Li, and Bradley A Malin. 2015. Ru policy frontiers for health data de-identification. *Journal of the American Medical Informatics Association*, 22(5):1029–1041.

A Appendices

A.1 Hyperparameters for the PHI NER

Parameters	Options
Linesearch	**MoreThuente**, StrongBacktracking, Backtracking
Max iterations	50, **100**, 150, 200, 250
Min freq	0, 3, **5**
Period	**5**, 10, 15
Num memories	3, 6, **9**, 12
c1	0.1, **0.05**, 0.01, 0.05, 0.001, 0,0005
c2	0.1, 0.05, **0.01**, 0.05, 0.001, 0,0005
epsilon	**1e-02**, 1,E-03, 1,E-04, 1,E-05, 1,E-06
delta	**1e-02**, 1,E-03, 1,E-04, 1,E-05, 1,E-06
transitions?	True, **False**
states?	**True**, False

Table 2: This is the options used for the hyper-parameter optimisation with random search. The bold ones are the parameters that together produced the best results for the 100 iterations.

A.2 Thresholds based on Grid Search

Main Non-PHI Threshold	Alt PHI Threshold
0.99999, 0.9999, 0.999, 0.99,	0.00001, 0.0001, 0.0005, 0.001,
0.95, 0.90, 0.85, 0.80,	0.005, 0.01, 0.05, 0.1,
0.75, 0.7, 0.6	0.2, 0.3, 0.4

Table 3: This table shows the options for the grid search used for choosing marginal thresholds for the different PHI models.

Optimised	Main Non-PHI Threshold	Alt PHI Threshold
F_1	0.75	0.1
F_4	0.99	0.05
F_{10}	0.999	0.001
F_{20}	0.9999	0.0001
F_{40}	0.99999	0.00001

Table 4: This table shows the best threshold for the marginal probability score for the predicted non-PHI label (Main Non-PHI Threshold) and the threshold for the next most probable label (Alt PHI Threshold) based on a grid search.