Deidentifying a Norwegian clinical corpus - An effort to create a privacy-preserving Norwegian large clinical language model

Phuong Dinh Ngo[1,2], **Miguel Tejedor**[1,3], **Therese Olsen Svenning**[1]
Taridzo Chomutare[1,4], **Andrius Budrionis**[1,2], **Hercules Dalianis**[1,5]
[1]Norwegian Centre for E-health Research, Tromsø, Norway
[2]Department of Physics and Technology, UiT The Arctic University of Norway
[3]Department of Mathematics and Statistics, UiT The Arctic University of Norway
[4]Department of Computer Sciences, UiT The Arctic University of Norway
[5]Department of Computer and Systems Science, Stockholm University, Sweden
Corresponding author: `Phuong.Dinh.Ngo@ehealthresearch.no`

Abstract

This study discusses the methods and challenges of deidentifying and pseudonymizing Norwegian clinical text for research purposes. The results of the NorDeid tool for deidentification and pseudonymization on different types of protected health information were evaluated and discussed, as well as the extension of its functionality with regular expressions to identify specific types of sensitive information. This research used a clinical corpus of adult patients treated in a gastro-surgical department in Norway, which contains approximately nine million clinical notes. The study also highlights the challenges posed by the unique language and clinical terminology of Norway and emphasizes the importance of protecting privacy and the need for customized approaches to meet legal and research requirements.

1 Introduction

Today with the European General Data Protection Regulation (GDPR) law, and the Norwegian law for the processing of personal information *Lov om behandling av personopplysninger (personopplysningsloven)* it is notoriously difficult to get access to electronic patient record texts to perform research.

First of all, one needs to submit an application to the *Norwegian Regional Committees for Medical and Health Research Ethics (REK)* and after that approval, one needs to ask *Personvernsombud (PVO)* at the local hospital to access the data. One way to make it easier is to process the data before using it for research by sanitising the data, that is to deidentify and then pseudonymise it, very similar to what has been carried out in (Vakili et al., 2022).

2 Related research

The field of deidentifying and pseudonymizing clinical text for research purposes has been a subject of extensive research, with much of the previous work based on shared tasks related to datasets such as *i2b2* (now n2c2). In (Stubbs and Uzuner, 2015), and most studies model deidentification as a named entity recognition (NER) task, (Nadkarni et al., 2011). Making these datasets available to researchers has facilitated a lot of progress on this task over the years; starting with traditional NLP methods (Stubbs et al., 2015), then with deep learning using word embeddings, (Dernoncourt et al., 2017), then more recently to deep learning methods using contextual embeddings or large language models, (Vakili and Dalianis, 2022). These benchmark datasets partially solved the need for standardised evaluation metrics to facilitate the comparison and improvement of different deidentification methods. Regarding the generation of pseudonyms or surrogates there is a nice description carried out by Olstad et al. (2023) where the authors elaborate on the replacement at different generalisation levels.

More recent research have shown promising results in the field. Among others, López-García et al. (2023) conducted a study on the automatic deidentification of medical documents in Spanish. The study developed two different deep learning-based methodologies for the task and also developed a data augmentation procedure to increase the number of texts used to train the models. Vakili et al. (2022) carried out deidentification and pseudonymisation of 17.9 Gb of Swedish clinical text using a Swedish clinical BERT model called SweDeClinBERT. The process of deidentification took over two weeks, while pseudonymisation, replacing the found entities with pseudonyms or surrogates, was ready in a couple of days since it is a rule based approach. In total 83,914,340 sensitive entities were found in 49,715,558 sentences encompassing 2.8 billion words.

Zheng et al. (2021) reviewed the recent research for ensuring the correct usage of regular expressions, which is crucial for identifying specific types of sensitive information.

However, there is a growing focus on addressing the challenges posed by the diverse and nuanced nature of clinical narratives, including variations in language use, context, and medical jargon. For instance, different institutions have different standards on how they treat their electronic health record narratives. This highlights the importance of documenting deidentification processes in diverse contexts, such as Norway in this case.

3 Methods and Materials

3.1 Methods

The methods used are a combination of deep learning methods and rule-based methods in the form of regular expressions.

The NorDeid deidentification and pseudonymisation tool was used in this study. NorDeid utilises the ScandiBERT[1] language model based on all Scandinavian languages and fine-tuned on the Swedish Stockholm EPR PHI Pseudo Corpus augmented with Danish and Norwegian personal names, (Lamproudis et al., 2023). ScandiBERT is a Bidirectional Encoder Representations (BERT) that was specifically fine-tuned for understanding and processing the Scandinavian languages, including Danish, Norwegian, and Swedish. NorDeid's functionality was extended with a number of regular expressions to identify *email-addresses, Norwegian social security numbers, user name* and *family numbers*, and used to identify the Protected Health Information (PHI) described in Table 1. PHIs are entities in a text that can reveal the identity of a person.

The chosen strategy to sanitise the text is first the NER identification of the PHIs and secondly to pseudonymise the found PHI by replacing them with similar surrogates, (Dalianis, 2019). For example: A last name is replaced with another random last name, the same name is replaced with the same random name to keep the coherence within the discourse. Female names are replaced with another random female name. A gender-neutral first name is replaced with another random gender-neutral first name. A location is replaced with another location nearby.

The *HIPS, Hidden in Plain Sight* strategy proposed by Carrell et al. (2019) was used in this study which implies removing the tags around the identified and pseudonymised PHIs so the PHIs that have been missed to be identified will be hidden among the pseudonymised PHIs.

PHI classes	**Found PHIs**
First Name	26,250,587
Last Name	29,793,462
Phone Number	14,227,411
Full Date	20,063,639
Date Part	19,866,503
Health Care Unit	84,232,994
Location	11,407,571
Organisation	5,292,142
Family Number	15,215,076
Social Security Number	700,527
Email	125,572
User name	4,126,831
Summary	227,179,610

Table 1: The table presents the PHI-classes[2] to be deidentified.

3.2 Materials

A clinical corpus called ClinCode Gastro Corpus containing 31,378 adult patients treated between the years 2017 to 2022 at the Gastro-Surgical department at the University Hospital of North Norway, Tromsø was used[3]. The dataset includes approximately 8.8 million clinical notes (in total, 27.6 Gb).

4 Application of method

A server *Republic of Gamers* with the operating system Debian Linux installed and equipped with two GPUs (ASUS Geforce RTX 3090), 64 Gb of internal memory (RAM) (2 x 32GB 3200 MHz DDR4), 8 TB Gen4 x4 M.2 NVMes SSD hard disc etc and not connected to the Internet was used

[1] https://huggingface.co/vesteinn/ScandiBERT

[2] The PHI class *Age* was at some point excluded from the execution of NorDeid after some discussion within the research group, since it was not considered sensitive, but it can easily be included again.

[3] This research was approved by The Norwegian Regional Committees for Medical and Health Research Ethics (REK) North, decision number 260972

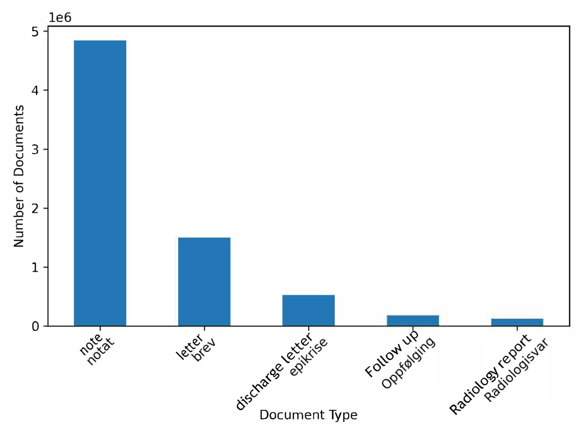


Figure 1: Top five types of clinical notes in the data. The letters are not clinical notes and will not be used in the research.

for the deidentification task. The server was also encrypted and situated in a server room only accessible to researchers who were specially authorised to work with the data and signed a confidentiality agreement. The server also remains offline during the project.

The process to deidentify and pseudonymise the corpus took approximately one week. The results can be seen in Table 2.

The evaluation of the current version of NorDeid is based on a comparison between human annotations and predictions made by the model. The evaluation dataset consists of 19 clinical notes that encompass about 13,000 tokens, annotated in the *CoNLL* format, a popular schema for text annotation used in natural language processing. The annotations target various entities of PHIs listed in Table 1.

The performance of the model is quantitatively measured using standard metrics: precision, recall, and F_1-score. Precision measures the proportion of correct positive identifications made by the model, recall assesses the model's ability to identify all relevant instances, and the F_1-score provides a harmonic mean of precision and recall, offering a balance between the two.

5 Analysis

The evaluation results are presented in a detailed format as shown in Table 2, covering various types of PHIs. The model shows varying levels of effectiveness across different PHI types. For example, it performs well in identifying entities like *First_Name, Full_Date*, and *Phone_Number*, but it struggles with *Family_Number, Organisation*, and *Social_Security_Number*.

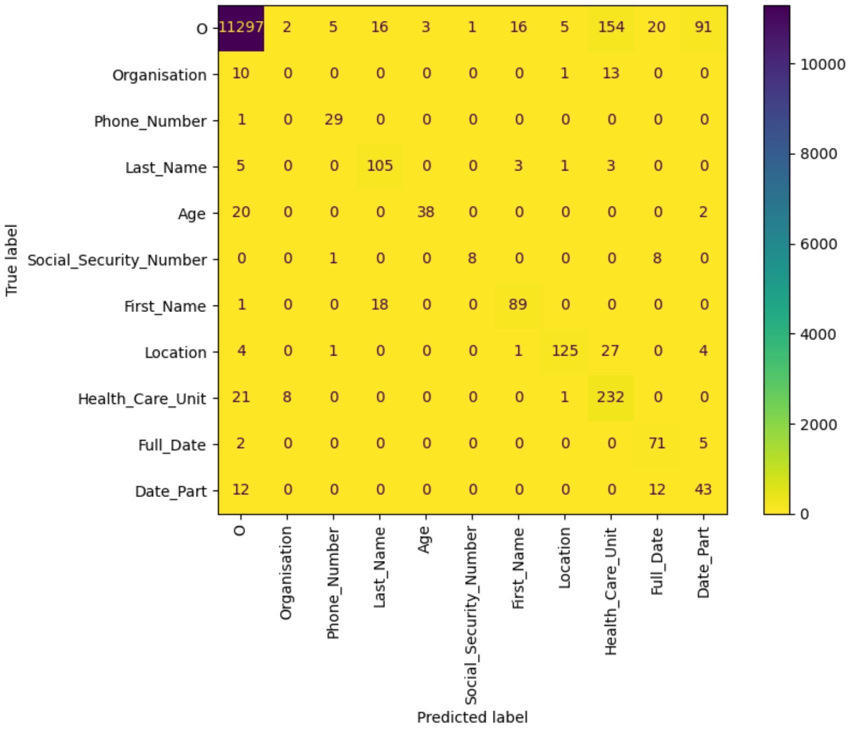
The model achieves its highest F_1-scores with *Full_Date* (0.76), *Phone_Number* (0.73), and *First_Name* (0.68). These results indicate a strong ability to recognise and accurately tag full dates and names in clinical notes. Although the model shows no capability in correctly classifying *Family_Number, Organisation*, and *Social_Security_Number*, NorDeid was able to identify those entities as PHIs, according to the confusion matrix (Figure 2). By looking at the average scores (micro average, macro average and weighted average), the model demonstrates moderate effectiveness with a weighted average F_1-score of 0.53. While this indicates potential utility in a clinical setting, there is notable room for improvement.

Figure 2 shows the entity confusion matrix. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. For PHI types such as *Health_Care_Unit*, *Full_Date*, *First_Name*, and *Last_Name*, there is a higher number of true positives. This shows a strong alignment with human annotations. Certain types of PHIs, such as *Organisation* and *Social_Security_Number*, have higher false positives and false negatives. This suggests the challenge in classifying these PHIs correctly. There are also a high number of misclassifications between *Location* and *Health_Care_Unit*, as well as between *Date_Part* and *Full_Date*. This could be due to the similarity in format and context between these types of PHIs.

PHI class	**Precision**	**Recall**	**F_1-score**
Age	0.30	0.32	0.31
First_Name	0.61	0.76	0.68
Last_Name	0.66	0.70	0.68
Full_Date	0.65	0.93	0.76
Date_Part	0.28	0.44	0.34
Health_Care_Unit	0.29	0.40	0.34
Location	0.75	0.63	0.68
Organisation	0.00	0.00	0.00
Phone_Number	0.60	0.92	0.73
Social_Security_Number	0.00	0.00	0.00
Family_Number	0.00	0.00	0.00
Username	0.00	0.00	0.00
micro avg	0.47	0.59	0.52
macro avg	0.34	0.43	0.38
weighted avg	0.49	0.59	0.53

Table 2: Evaluation results of NorDeid on 19 random clinical notes, approximately 13,000 tokens

Precision, recall, and F_1-score do not consider



True label \ Predicted label	O	Organisation	Phone_Number	Last_Name	Age	Social_Security_Number	First_Name	Location	Health_Care_Unit	Full_Date	Date_Part
O	11297	2	5	16	3	1	16	5	154	20	91
Organisation	10	0	0	0	0	0	0	1	13	0	0
Phone_Number	1	0	29	0	0	0	0	0	0	0	0
Last_Name	5	0	0	105	0	0	3	1	3	0	0
Age	20	0	0	0	38	0	0	0	0	0	2
Social_Security_Number	0	0	1	0	0	8	0	0	0	8	0
First_Name	1	0	0	18	0	0	89	0	0	0	0
Location	4	0	1	0	0	0	1	125	27	0	4
Health_Care_Unit	21	8	0	0	0	0	0	1	232	0	0
Full_Date	2	0	0	0	0	0	0	0	0	71	5
Date_Part	12	0	0	0	0	0	0	0	0	12	43

Figure 2: Entity level confusion matrix

the expected chance agreements that occur when humans annotate instances. We calculated the inter-annotator agreement to measure how well two different annotators made the same annotation decision. The two independent annotators annotated the same subset of clinical notes following the annotation guidelines developed in this work to qualitatively validate the labels. The inter-annotator agreement calculated using Cohen's Kappa was 0.86, indicating almost perfect agreement, (Landis and Koch, 1977).

6 Challenges

A major challenge was encountered in the beginning when trying to identify and classify sensitive data. This required a detailed understanding of the type and extent of sensitive information applicable to Norwegian texts, which included a wide variety of personal identifiers and confidential medical details. The difficulty was further increased by the subtle differences in language and clinical terminology that are unique to Norway. An example of illustrating this challenge is *personnummer*, which refers to a Norwegian social security number. This number is highly sensitive since it is unique to each person and contains information such as date of birth and gender. The model needed to distinguish between actual *personnummer* instances and other similar-looking numerical sequences. There are also different ways to represent this number depending on how healthcare professionals annotate notes. For example, a *personnummer* of *01010112345* can be rewritten as *010101-12345* or *010101 12345* or *01 jan 01 12345*.

The task of data management, such as cleaning and formatting, can be challenging. This included ensuring the data was in the correct text format, linking the data correctly, and dealing with issues when tokenizing Norwegian clinical texts. These processes were essential to ensure the data was accurate and reliable before using them with the model. Annotation of clinical texts requires a lot of resources. This process required both time and expertise, particularly in the medical domain. Therefore, providing enough resources for annotation is a major challenge in ensuring the overall efficiency and precision of the deidentification procedure.

Implementing the model to deidentify Norwegian clinical texts is also computationally intensive. It requires substantial computational resources, including processing power and memory, which was a limiting factor. Operating in an offline environment also introduced additional constraints, particularly in setting up the environment for model training and debugging code. This scenario limits the ability to take advantage of cloud computing resources and requires reliance on local computational capabilities.

Finally, there is a potential challenge in mitigating biases in the training and the output produced by NorDeid. Since the model relies on existing datasets for training, there is a risk of carrying the biases that exist within these datasets. Therefore, ensuring the unbiased and equitable functioning of the model in the deidentification of Norwegian clinical texts is essential and should not be overlooked.

7 Discussion

Each deidentification system needs to be customised on which PHIs to remove depending on the research task and type of data or what each country has for laws or rules. Lawyers and physicians do not always agree on which PHI is sensitive. For example, *Health Care Unit* can be valuable to keep sometimes, *Age* can also be important in certain research, most clinical researchers want to keep the class name while computer scientists consider it is much safer to replace identified PHIs with pseudonyms or surrogates, (Vakili and Dalianis, 2022). In the example with *Age* it can be replaced with an age close to the actual age, for example, random ± 2-3 years.

8 Conclusion

In conclusion, this paper has discussed the challenges and methods involved in deidentifying and pseudonymizing Norwegian clinical text for research purposes. The use of the NorDeid tool and regular expressions for identifying specific types of sensitive information proved effective in the deidentification process. The research highlighted the importance of privacy preservation and the need for tailored approaches to meet legal and research requirements. The significance of mitigating potential biases in the training and output of deidentification models were also emphasized.

9 Future work

The plan for the produced pseudonymised gastro corpus now called *ClinCode Gastro Pseudo Corpus* is to create a Norwegian Clinical BERT Model using the publicly available Norwegian language model NorBERT[4] based on general Norwegian Bokmål and Nynorsk, and perform continued pretraining from NorBERT on the pseudonymised gastro corpus. The aim of this is twofold first to improve the deidentification tool NorDeid and secondly to make a privacy preserved Norwegian large clinical language model available to researchers worldwide and improve the result of the current Norwegian clinical text mining. We will also extend the ClinCode Gastro Corpus with more manually annotated PHIs improve the performance of the NorDeid tool. The NordDeid tool is available for use by other researchers and research groups.

10 Limitations

The study may be limited by the availability of annotated Norwegian datasets for training and evaluating deidentification models. In addition to the limitation posed by the Norwegian language, it is important to note that there exist minor languages in Norway that were not considered in this study. The performance of the model has not been evaluated for these languages, which may introduce a bias in the results. This highlights a potential limitation of the study and underscores the importance of considering linguistic diversity in future research to ensure inclusivity and avoid bias.

The model's performance was not perfect and it had problems in classifying the PHIs in the correct PHIs class. In some cases, only parts of the health care unit or the social security number were identified, which led that only parts of it were pseudonymised, but NorDeid did its task in deidentifying and pseudonymising sensitive information. NorDeid also identified some false positives such as parts of ICD-10 codes or Drug names (as last names). In the Appendix some examples are shown. In the examples the SGML tags are left.

The clinical text used in the study was extracted only from a gastro-surgical department. Therefore, there may be a potential lack of generalizability of the findings to other healthcare domains or organizations. Finally, the potential impact of biases that exist in training data and how they affect deidentification and pseudonymization needs further investigation.

Acknowledgements

The research was funded by the Norwegian Research Council under the project *ClinCode - Computer-Assisted Clinical ICD-10 Coding for improving efficiency and quality in healthcare*.

[4] NorBERT, http://wiki.nlpl.eu/Vectors/norlm/norbert.

References

David S Carrell, David J Cronkite, Muqun Li, Steve Nyemba, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. 2019. The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight. *Journal of the American Medical Informatics Association*, 26(12):1536–1544.

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

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*.

Anastasios Lamproudis, Sara Mora, Therese Olsen Svenning, Torbjørn Torsvik, Taridzo Chomutare, Phuong Dinh Ngo, and Hercules Dalianis. 2023. De-identifying Norwegian Clinical Text using Resources from Swedish and Danish. In *AMIA Annual Symposium Proceedings*, volume 2023. American Medical Informatics Association.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Guillermo López-García, Francisco J. Moreno-Barea, Mesa Héctor, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. Named Entity Recognition for De-identifying Real-World Health Records in Spanish. In *Computational Science – ICCS*, pages 228–242.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of Replacement Options in Text Sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 292–300.

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.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

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.

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.

Li-Xiao Zheng, Shuai Ma, Zu-Xi Chen, and Xiang-Yu Luo. 2021. Ensuring the Correctness of Regular Expressions: A Review. *International Journal of Automation and Computing*, 18:521–535.

Appendix: De-identified clinical texts

Here follows examples with de-identified and pseudonymised Norwegian clinical text where the SGML tags has been kept (hence not HIPS), for pedagogical purposes.

1. Var henvist av egen lege til en kolonskopi som skulle vært tatt den <Date_Part>28.</Date_Part> < Date_Part>08,</Date_Part> men har utsatt denne grunnet operasjon.

2. Gjenomgikk så <Full_Date>17.02.19</Full_Date> fedmekirurgi (type Gastric sleeve) ved <Health_Care_Unit>en vårdenhet</Health_Care_Unit>, reoperert <Date_Part>20.02</Date_Part> pga blødning ved <Health_Care_Unit>en vårdenhet</Health_Care_Unit>.

3. Godkjent av/skrevet av lege i spesialisering 2 <First_Name>Signe</First_Name> <Last_Name>Rybakk</Last_Name> / <User_Name>/Sry001</User_Name> Da hun bor i <Location>Horten</Location> ble hun sendt hjem og kommer derfor i dag for kontroll.

4. <Full_Date>26.01</Full_Date> 17 Journalnotat SO, <Health_Care_Unit>en vårdenhet</Health_Care_Unit> <Health_Care_Unit>Bergen</Health_Care_Unit> v/Overlege endokrinologi <First_Name>Carrie</First_Name> <Last_Name>Rammus</Last_Name> /Cra2377aaa Pasienten er overflyttet hit fra <Location>Kongsberg</Location> pga akutt nekrotiserende pancreatitt med påfølgende langt behandlingsforløp og intensivt opphold.

5. J<Date_Part>UG05</Date_Part> Rektoskopi md biopsi

In Example 1. above, one can observe that the de-identifier splitted the *Date_Part* in two parts while *28.08* should be encompassed by one *Date_Part* tag.

In Example 4. the de-identifier missed to tag number *17* as in year *2017*, while *26.01* was tagged as *Full_Date* while it is actually a *Date_Part*. To be correct *26.01 17* should be tagged as *Full_Date*, moreover in the same example the User name */Msø2377* is not tagged.

In Example 5. a part of a procedure code is wrongly identified as a *Date_Part*.