Automatic Annotation of Training Data for Deep Learning Based De-identification of Narrative Clinical Text

Martin Sundahl Laursen[1], Jannik Skyttegaard Pedersen[1], Pernille Just Vinholt[2] and Thiusius Rajeeth Savarimuthu[1]

[1]The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Denmark
[2]Department of Clinical Biochemistry, Odense University Hospital, Denmark

Abstract
Electronic health records contain information about patients' medical history which is important for research but the text must be de-identified before use. This study utilized dictionaries constructed from publicly available lists of identifiers to automatically annotate a training dataset for a named entity recognition model to de-identify names, streets, and locations in Danish narrative clinical text. Ambiguous identifiers were not annotated if they occurred more than expected for an identifier. The model had recall 93.43%, precision 86.10%, and F1 89.62%. We found that the model generalized from the training data to achieve better performance than simply using the dictionaries to directly annotate text.

Keywords
de-identification, electronic health records, named entity recognition, automatic annotation, deep learning

1. Introduction

Electronic health records (EHR) contain information about patients' contact with the healthcare system including important information about medical history, e.g. symptoms, diagnoses, and treatments. Diagnoses are also registered using International Classification of Diseases 10 codes for administrative purposes. However, not all relevant patient information is represented in codes, e.g symptoms. Further, codes are often incorrect [1, 2, 3, 4, 5] and can therefore not replace the narrative clinical text in EHRs as a source of information.

Apart from treatment of patients, the EHR data are important for e.g. research and education but as they contain personally identifiable information, explicit consent from the affected individual must be given, or the data must be de-identified before being used for secondary purposes [6, 7]. The US Health Insurance Portability and Accountability Act (HIPAA) defines which identifiers must be removed according to the Safe Harbor method for de-identification[1].

WNLPe-Health 2022, December 18, 2022, Delhi, India
msla@mmmi.sdu.dk (M. S. Laursen); jasp@mmmi.sdu.dk (J. S. Pedersen); pernille.vinholt@rsyd.dk (P. J. Vinholt); trs@mmmi.sdu.dk (T. R. Savarimuthu)
0000-0001-5684-1325 (M. S. Laursen); 0000-0002-7066-1563 (J. S. Pedersen); 0000-0002-2035-0169 (P. J. Vinholt); 0000-0002-2478-8694 (T. R. Savarimuthu)

[1]Guide available at https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

The identifiers include, among others, names, street addresses, and locations including city, county, and precinct.

Manual de-identification is a time consuming task and, therefore, large datasets are impractical and expensive to de-identify manually. Natural language processing techniques for automatic de-identification may alleviate this task.

This study utilizes dictionaries of identifiers and a novel way of dealing with ambiguous identifiers to automatically annotate a training dataset for a named entity recognition (NER) model to de-identify names, streets, and locations in Danish narrative clinical text.

A method for automatic annotation of training datasets is useful for developing de-identification deep learning models for low-resource languages like Danish where annotated datasets and trained models for de-identification of specific identifier types are not always publicly available.

The main contributions of this paper are:

- We train a NER model to de-identify names, streets, and locations in Danish narrative clinical text with recall 93.43%, precision 86.10%, and F1 89.62%.
- We use dictionary-based automatic annotation of training data for the NER model utilizing our novel method for annotation of ambiguous identifiers guided by occurrence rates in the text and population.
- We find that the NER model can generalize from the dataset to achieve better performance than simply using the dictionaries to directly annotate text.

2. Related Work

Previous studies on automatic de-identification of narrative clinical text used rule-based methods, machine learning methods, and hybrid methods combining both. We found no studies that, similar to ours, used automatic annotation of data for training a machine learning model to de-identify narrative clinical text.

In studies that used rule-based methods, pattern matching or dictionaries were used to search for identifiers in the text [8, 9, 10, 11, 12, 13]. Rule-based methods rely on domain experts to define the rules and it is difficult to cover all cases. They generally cannot distinguish ambiguous identifiers, i.e. words that can both be an identifier and a non-identifier depending on the context. Pantazos et al. [12] de-identified Danish text using dictionaries with an F1 score of 95.7% on a random sample of 369 EHRs. They identified ambiguous identifiers by matching identifiers to a database of non-identifiers. As they replaced identifiers with pseudo-identifiers, their approach to ambiguous identifiers was to delete the record unless the identifier appeared more than 200 times to not disclose their replacement rule.

Studies that used machine learning methods mainly used recurrent neural networks, conditional random fields, and combinations of the two [14, 15, 16, 17, 18, 19]. All studies that utilized machine learning used a manually annotated dataset for training the models. Machine learning methods and in particular deep learning architectures such as Long Short-Term Memory [20] and transformer [21] networks are able to distinguish ambiguous identifiers based on the context of the whole sentence. Some recent studies have used transformer networks for automatic de-identification of narrative clinical text [22, 19, 23, 24, 25, 26]. A disadvantage of machine

learning methods is their need for a large expert-annotated training dataset specific to the domain.

Finally, the studies that were most similar to ours used hybrid methods, combining rule-based and machine learning methods in ensembles or pipelines to improve the annotation workload and model performance [27, 28, 29, 30, 25]. Two studies used rule-based methods in other ways than for directly classifying identifiers. McMurry et al. [27] used pattern matching to contribute part of a feature set for classification by a machine learning model which resulted in a F1 score of 76% on a custom test set of 220 discharge summaries. Jian et al. [28] used pattern matching to create a dense corpus of identifiers for manual annotation before being input to a machine learning model. It had an F1 score of 94.6% when cross-validating on 3,000 clinical documents.

3. Methods

In this paper, we first constructed lists of name, street, and location identifiers. We compared the identifiers to a database of non-identifying words to determine which identifiers were ambiguous—e.g. the name 'Hans' is ambiguous because it is also a pronoun (Danish for 'his'). This dictionary-based method is similar to that of e.g. Pantazos et al. [12] except in this paper, we used it to annotate training data for a deep learning model instead of using direct dictionary-based de-identification. Additionally, we utilized a novel method for annotation of ambiguous identifiers and tested different ceiling values above which words were removed from the list of identifiers if they occured in the text at a higher rate than would be expected for an identifier. We searched for and annotated identifiers in Danish narrative clinical text and constructed a training set of sentences with no or only unambiguous identifiers. Finally, the training set was used to train a NER de-identification model. The goal was for the model to generalize from the training samples with no or only unambiguous identifiers to also correctly classify ambiguous identifiers. This process is detailed in the rest of this section. We make our code publicly available[2].

3.1. Data

3.1.1. Corpus

We extracted 150,000 random sentences with a length between 8 and 70 words from EHRs from Odense University Hospital between 2015 and 2020. Sentences were lowercased and tokenized, and consecutive underscores and hyphens were reduced to a single instance.

3.1.2. Identifiers

The identifier types were names, streets, and locations. Locations included cities, municipalities, regions, and provinces.

For the name identifiers, we obtained lists of all male first names, female first names, and last names in the Danish population as of January 2021 from Statistics Denmark.

[2]https://github.com/jannikskytt/clinical_de-identification

For the street identifiers, we used a database of all Danish addresses from the Address Web Services of the Agency for Data Supply and Efficiency of Denmark[3]. Each address included street name, addressing street name (could be identical to street name), city name, potential supplemental city name, municipality, region, and province. From the database, a list of unique street names including addressing street names was constructed.

For the location identifiers, we used the same database of all Danish addresses. A list of unique locations including city names, supplemental city names, municipalities, regions, and provinces was constructed.

Data cleaning consisted of lowercasing and removing single-letter and empty and corrupted identifiers including various placeholders.

A rate of occurrence in the Danish population was calculated for each identifier by dividing their occurrence in the population by the sum of all occurrences for that identifier type. For each identifier type, duplicates were merged by adding the rates of occurrence.

3.1.3. Non-identifiers

Non-identifiers were words that in their context did not identify names, streets, or locations. Such words included both common general domain words and specialized words from the clinical domain such as symptoms, diseases, and treatments. The database of non-identifiers was constructed from multiple text sources from the general and clinical domains which did not contain any of the three identifier types.

The text sources were:

- The Danish orthographic dictionary containing all Danish words, their conjugations, and abbreviations [31].
- Product names from the list of authorized medicinal products in Denmark[4].
- Medical abbreviations collected from different electronic sources (Appendix A)
- All term entries in the Description tables of the SNOMED CT vocabulary of clinical terminology (international version with Danish extension).
- The Danish healthcare system's classification system for symptoms, diagnoses, and operations[5].

3.2. Ambiguous Identifiers

An identifier could be ambiguous for two reasons. One reason was that it had multiple different identifier types, e.g. 'Kolding' is both a location and a name. In that case, the identifiers' rates of occurrence were added. Another reason was that it was also a non-identifier. To find those cases, identifiers were matched against the database of non-identifiers using a regular expression that ignored case (regex). If an identifier was matched to a non-identifier, it was ambiguous.

[3]All datasets were downloaded from https://download.aws.dk/
[4]Available at https://laegemiddelstyrelsen.dk/en/
[5]Available at https://sundhedsdatastyrelsen.dk/

3.3. Automatic Annotation

For the automatic annotation of identifiers in sentences, specifically dealing with ambiguous identifiers, we introduced a measure for the likelihood of a word being a non-identifier vs. identifier for the specific corpus. The measure was the ratio between the rate with which the word occurred in sentences in the corpus as either identifier or non-identifier, and the rate of occurrence in the Danish population as identifier: $ratio = r_{corpus}/r_{population}$. A ratio above 1 meant that the word had a higher rate of occurrence in the corpus than as an identifier in the Danish population. This could indicate that it in most cases occurred as a non-identifier in the corpus. A ratio below 1 could indicate that the word in most cases occurred as an identifier.

The rate of occurrence in the corpus was calculated for each identifier by searching through all sentences using a regex, counting the number of occurrences, and dividing by the total number of sentences. The ratio was then calculated using the equation.

Next, a regex was used to search for and annotate identifiers in the sentences. Words that were unambiguous identifiers were annotated with their single identifier type. Words that were ambiguous because they had multiple identifier types were annotated with both. Words that were ambiguous because they were both an identifier and a non-identifier were annotated with their identifier type and a non-identifier tag with two exceptions: (1) if the ratio was below 1, they were annotated only with their identifier type, and (2) if the ratio was above a set ratio ceiling, the identifier was not annotated, i.e. kept as a non-identifying word.

Finally, all annotated sentences were postprocessed in the following order:

1. If an ambiguous identifier was the same type as a neighbor identifier, it was converted to that type.
2. If a single letter was between two name identifiers, it was taken as a middle initial and converted to a name identifier.
3. Identifiers of the same type which were next to each other were converted to a single identifier consisting of multiple words.

We tested values for the ratio ceiling on a binary logarithmic scale from 1 to 262,144.

3.4. Named Entity Recognition Model

We used the automatically annotated sentences to create multiple datasets, based on different values for the ratio ceiling, for training Princeton University Relation Extraction system (PURE) [32] NER models to de-identify name, street, and location identifiers in the corpus.

3.4.1. Datasets

The validation and test sets each contained 1,500 sentences. They were annotated for names, streets, and locations by one of the authors using the CLAMP software [33]. The sentences for the validation and test sets were selected by setting the ratio ceiling to the median ratio of all identifiers and choosing 500 sentences with no identifiers, 500 with only unambiguous identifiers, and 500 with at least one ambiguous identifier. The distributions of types of ambiguous and unambiguous identifiers were approximately the same as in the entire corpus. Selecting the

validation and test sets in this way ensured that as many models as possible would experience varying sentences with regards to types, ambiguity, and number of identifiers.

While the validation and test sets were human annotated and fixed for all models, the training sets were annotated automatically using the described method and varied with each of the tested ratio ceilings used for the automatic annotation. Training sets were constructed from all sentences not used for the validation and test sets. Only sentences with no or unambiguous identifiers were selected for the training sets since the NER model was only trained with unambiguous samples. In cases where the number of sentences containing no identifiers was higher than the number containing identifiers, the former was downsampled to the latter.

All datasets were converted to the structure used by PURE.

3.4.2. Training of Model

For each training set automatically annotated with the different ratio ceilings, a PURE NER model was trained with a publicly available uncased Danish pretrained BERT [34] model[6] as base. The default hyperparameters of PURE were used (see [32]) except a context window of 0. Models were trained until convergence (maximum 100 epochs). The F1 score on the validation set was used to select the best model checkpoint from each training.

3.4.3. Evaluation of Model

The best performing model on the current data was found by evaluating the F1 scores on the test set. Performance on the three identifier types was evaluated in a confusion matrix. Additionally, for each model, we compared its test set performance to that of the dictionary-based method used for annotating its training set to see if the model generalized from its training data to improve performance.

The ratio ceiling used for automatic annotation of the training set for the best performing model was tested for model training with less available data to evaluate the minimum amount needed for top model performance.

Finally, we analyzed the effect of lowering the ratio ceiling to produce more training samples when there was less data than the minimum amount needed for top model performance.

4. Results

4.1. Identifiers

The list of identifiers had 449,997 unambiguous identifiers: 397,348 names, 48,859 streets, and 3,790 locations. 18,057 identifiers were ambiguous: 16,582 had a name type, 2,505 a street type, and 3,133 a location type. 3,859 of the ambiguous identifiers had more than one identifier type. 7,148 ambiguous identifiers matched a non-identifier in the Danish orthographic dictionary, 312 in authorized medicinal products, 406 in medical abbreviations, 9,890 in SNOMED CT, and 2,013 in the healthcare system's classification system. Identifiers had rates of occurrence in the population between 8.58e-06% and 30.39% with median 1.72e-05%.

[6]Available at https://github.com/certainlyio/nordic_bert

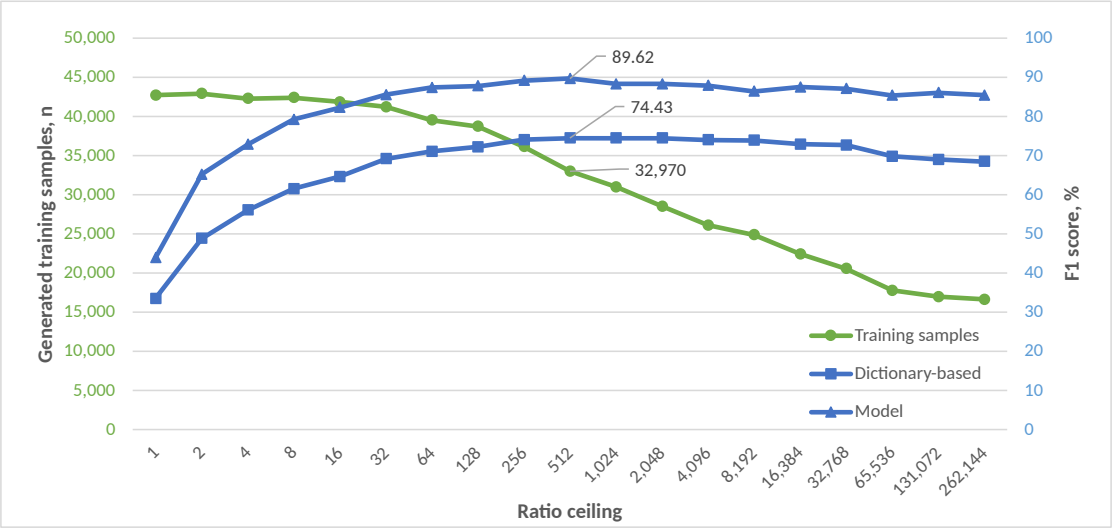

Figure 1: Model F1 and dictionary-based F1 in blue, right axis. Amount of samples in the training set in green, left axis.

Automatically annotated identifiers had corpus vs. population ratios between 4.42e-03 and 9.82e+06 (median 306.73). The highest ratio was 'københavn' (ambiguous location) while the lowest ratio was 'og' (ambiguous name and conjunction 'and').

4.2. Named Entity Recognition Model

Table 1
Identifiers in the validation and test sets.

	Validation set	Test set
Name	943	965
Street	77	97
Location	353	324
Total	1,373	1,386

Table 1 shows the distribution of identifier types in the human annotated validation and test sets.

Figure 1 shows the test set F1 scores for the models which training sets were automatically annotated with different ratio ceilings. The F1 score of the dictionary-based method and the amount of samples in the training sets are also plotted (further details in Appendix B).

The best model was trained with data automatically annotated with a 512 ratio ceiling (32,970 training set samples). It had a recall of 93.43%, a precision of 86.10% and a F1 of 89.62%. There was an upwards trend in the 1–512 ratio ceilings and downwards in 512–262,144. All model F1 scores were higher than the corresponding dictionary-based F1 scores. Most training set samples were produced with the ratio ceiling at 2 (42,894) while the least were produced by ratio ceiling 262,144 (16,614). Training time for the best model was 2 hours 4 minutes for 20

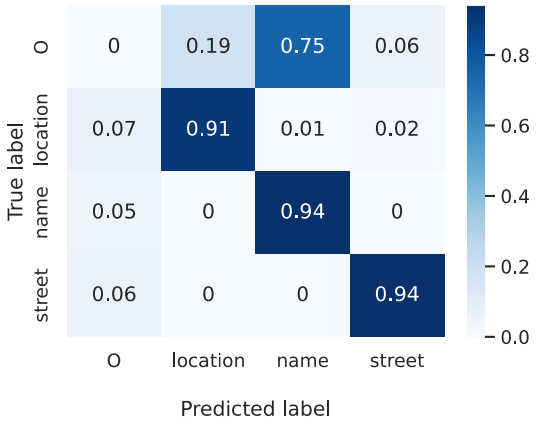

Figure 2: Test set confusion matrix. 'O' is non-identifiers for which only errors were counted.

epochs on a Nvidia Tesla v100 GPU.

Figure 2 shows the confusion matrix for model performance. 94% of street and name identifiers, and 91% of location identifiers were classified correctly. Non-identifiers were most often misclassified as names (75% of misclassifications).

Comparing test set performance to the dictionary-based method, the model correctly classified 283 identifiers that the dictionary-based method misclassified. The dictionary-based method correctly classified 13 identifiers that the model misclassified. 70 identifiers were misclassified by both the model and the dictionary-based method. Appendix C shows the performance of the model and the dictionary-based method on words that occurred in the test set both as non-identifiers and identifiers. E.g., for the word 'per', the model correctly classified it as an identifier (name) in 100% of cases and as a non-identifier (preposition: 'per') in 91% of cases. For the dictionary-based method, it was 57% and 100%, respectively. Note that the dictionary-based method could classify the same word differently because of the postprocessing steps where an ambiguous identifier could be converted to an unambiguous identifier under certain conditions. Among all words that occurred both as non-identifiers and identifiers, the model classified 92% of non-identifiers and 84% of identifiers correctly. For the dictionary-based method, it was 96% and 50%, respectively.

4.3. Analysis of Ratio Ceiling

We analyzed the effect of lowering the ratio ceiling to produce more training samples when there was less data than needed for top model performance.

The best performing model was trained on data automatically annotated with ratio ceiling 512 and had 147,000 sentences available from which 32,970 sentences were used for the training set. We lowered the amount of available data for automatic annotation with ratio ceiling 512 from 147,000 through to 12,000 sentences without any reduction in performance.

Next, we tested the effects on which ratio ceiling was the best when lowering the amount of available data below 12,000 sentences. We included ratio ceilings between 512 and 16 since they

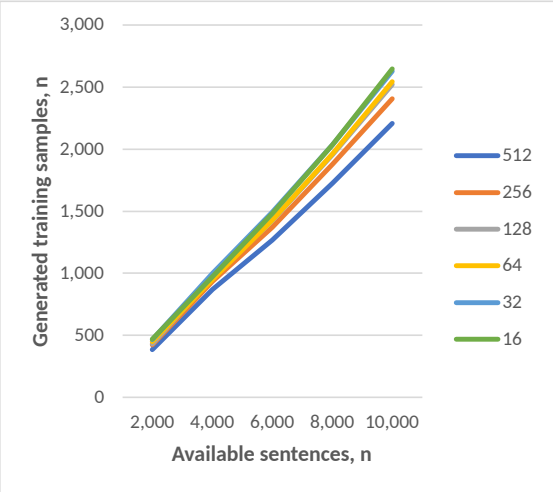

Figure 3: Amount of training set samples generated from the available data by automatically annotating with ratio ceilings 16–512.

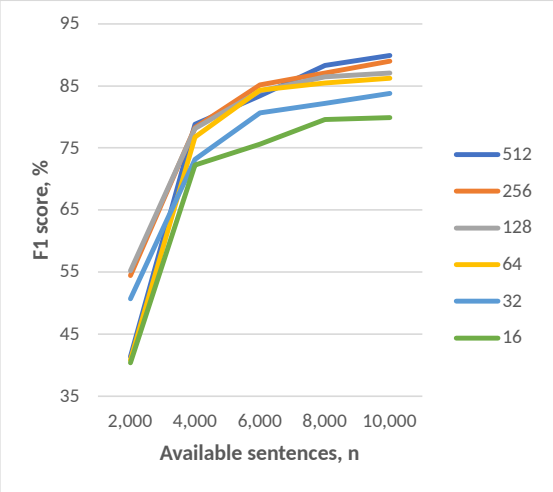

Figure 4: F1 scores for models with training sets automatically annotated with ratio ceilings 16–512 by amount of available data.

generated increasingly more samples for the training set (Figure 3). When the available data was less than 8,000 sentences, performance with lower ratio ceilings surpassed that of the 512 ratio ceiling in some cases (Figure 4).

5. Discussion

We used an automatically annotated training set to train a PURE NER deep learning model to de-identify names, streets and locations in Danish narrative clinical text with a recall of 93.43%, precision of 86.10%, and F1 score of 89.62%. Non-identifiers were most often misclassified as

names which may be caused by greater variability than for streets and locations.

We took a similar approach as Pantazos et al. [12] to identify ambiguous identifiers through matching of identifiers to a database of non-identifiers. While, for de-identification, they deleted records of ambiguous identifiers that occurred less than 200 times, we trained a deep learning model from an automatically annotated training set to de-identify ambiguous identifiers. For the automatic annotation, we handled ambiguous identifiers by calculating the ratio between the rate of occurrence in the corpus and the rate of occurrence in the population for every identifier. This method allowed an individual assessment if they should be annotated as an identifier or not in the training data—increasing the chance of model generalization. The ratio ceiling also allowed to balance the quality and amount of training data. Analyzing the ratio ceiling, we found that when less than 8,000 sentences were available, the extra samples provided by a lower ratio ceiling became more important than using the ratio ceiling that gave the highest quality of the training data. Lower ratio ceilings produced more training data because more ambiguous identifiers were considered non-identifiers resulting in fewer ambiguous sentences that had to be discarded from the training set.

We saw an increase in F1 from dictionary-based de-identification to annotating a training set with the dictionary-based method, training a NER model, and de-identifying with the trained model. This showed that the model generalized from the training data to better classify the ambiguous identifiers that the dictionary-based approach could not differentiate, and achieve better performance than simply using the dictionaries to directly annotate text. This is supported by the model correctly de-identifying 84% of words that occurred in the test set both as identifier and non-identifier. Only 50% of these words were de-identified by the dictionary-based method.

5.1. Limitations

It is a limitation to the study that the data came only from Odense University Hospital but the ratios were calculated using the rate of occurrence in the entire population of Denmark.

Future work includes de-identification of the rest of the HIPAA Safe Harbor identifiers since there is no guarantee that the presented methods will generalize to other identifiers. Since this study used lowercased data because only a lowercased Danish BERT base was available, exploring performance when keeping the case of training data is also part of future work.

6. Conclusions

We trained a NER deep learning model using automatically annotated data to de-identify names, streets, and locations in Danish narrative clinical text with recall 93.43%, precision 86.10%, and F1 89.62%. A model trained on data annotated with a dictionary-based method can generalize and surpass the performance of the dictionary-based method. A ratio ceiling of 512 works best for Danish narrative clinical text when more than 8,000 sentences are available.

The automatic de-identification method presented in this study can be adapted to all languages and domains if lists of identifiers and non-identifiers are available. Apart from the lists, the method does not need any external data as the input data to the de-identification model is used to train the model itself. This makes the method particularly useful for low-resource languages

where annotated datasets and trained models for de-identification of specific identifier types are not always publicly available.

References

[1] V. E. Valkhoff, P. M. Coloma, G. M. Masclee, R. Gini, F. Innocenti, F. Lapi, M. Molokhia, M. Mosseveld, M. S. Nielsson, M. Schuemie, F. Thiessard, J. van der Lei, M. C. Sturkenboom, G. Trifirò, Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk, Journal of Clinical Epidemiology 67 (2014) 921–931. URL: https://www.sciencedirect.com/science/article/pii/S0895435614000845. doi:https://doi.org/10.1016/j.jclinepi.2014.02.020.

[2] L. R. Øie, M. A. Madsbu, C. Giannadakis, A. Vorhaug, H. Jensberg, Ø. Salvesen, S. Gulati, Validation of intracranial hemorrhage in the norwegian patient registry, Brain and Behavior 8 (2018) e00900. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/brb3.900. doi:https://doi.org/10.1002/brb3.900. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.900.

[3] J. Delekta, S. M. Hansen, K. S. AlZuhairi, C. S. Bork, A. M. Joensen, The validity of the diagnosis of heart failure (I50.0-I50.9) in the danish national patient register, Dan Med J 65 (2018).

[4] T. L. Higgins, A. Deshpande, M. D. Zilberberg, P. K. Lindenauer, P. B. Imrey, P.-C. Yu, S. D. Haessler, S. S. Richter, M. B. Rothberg, Assessment of the Accuracy of Using ICD-9 Diagnosis Codes to Identify Pneumonia Etiology in Patients Hospitalized With Pneumonia, JAMA Network Open 3 (2020) e207750–e207750. URL: https://doi.org/10.1001/jamanetworkopen.2020.7750. doi:10.1001/jamanetworkopen.2020.7750. arXiv:https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2768537/higgins_2020_oi

[5] N. Wabe, L. Li, R. Lindeman, J. J. Post, M. R. Dahm, J. Li, J. I. Westbrook, A. Georgiou, Evaluation of the accuracy of diagnostic coding for influenza compared to laboratory results: the availability of test results before hospital discharge facilitates improved coding accuracy, BMC Medical Informatics and Decision Making 21 (2021) 168. URL: https://doi.org/10.1186/s12911-021-01531-9. doi:10.1186/s12911-021-01531-9.

[6] GDPR, Regulation (eu) 2016/679 (general data protection regulation), ???? URL: https://gdpr-info.eu/.

[7] HIPAA, Health insurance portability and accountability act of 1996 (hipaa), public law 104-191, ???? URL: https://www.hhs.gov/hipaa/for-professionals/index.html.

[8] B. A. Beckwith, R. Mahaadevan, U. J. Balis, F. Kuo, Development and evaluation of an open source software tool for deidentification of pathology reports, BMC Medical Informatics and Decision Making 6 (2006) 12.

[9] I. Neamatullah, M. M. Douglass, L.-W. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, G. D. Clifford, Automated de-identification of free-text medical records, BMC Medical Informatics and Decision Making 8 (2008) 32.

[10] E. Chazard, C. Mouret, G. Ficheur, A. Schaffar, J.-B. Beuscart, R. Beuscart, Proposal and evaluation of fasdim, a fast and simple de-identification method for unstructured free-

text clinical records, International Journal of Medical Informatics 83 (2014) 303–312. URL: https://www.sciencedirect.com/science/article/pii/S1386505613002463. doi:https://doi.org/10.1016/j.ijmedinf.2013.11.005.

[11] S. Y. C. H. J. P. J. L. Y. L. M.-S. C. C.-M. K. W.-S. L. J. H. Shin Soo-Yong, Park Yu Rang, A de-identification method for bilingual clinical texts of various note types, jkms 30 (2015) 7–15. URL: http://www.e-sciencecentral.org/articles/?scid=1022920. doi:10.3346/jkms.2015.30.1.7. arXiv:http://www.e-sciencecentral.org/articles/?scid=1022920.

[12] K. Pantazos, S. Lauesen, S. Lippert, Preserving medical correctness, readability and consistency in de-identified health records, Health Informatics Journal 23 (2017) 291–303. URL: https://doi.org/10.1177/1460458216647760. doi:10.1177/1460458216647760. arXiv:https://doi.org/10.1177/1460458216647760.

[13] V. Menger, F. Scheepers, L. M. van Wijk, M. Spruit, Deduce: A pattern matching method for automatic de-identification of dutch medical text, Telematics and Informatics 35 (2018) 727–736. URL: https://www.sciencedirect.com/science/article/pii/S0736585316307365. doi:https://doi.org/10.1016/j.tele.2017.08.002.

[14] H. Fabregat, A. Duque, J. Martinez-Romo, L. Araujo, De-identification through named entity recognition for medical document anonymization, Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019).

[15] K. Kajiyama, H. Horiguchi, T. Okumura, M. Morita, Y. Kano, De-identifying free text of japanese electronic health records, Journal of Biomedical Semantics 11 (2020) 11.

[16] L. Lange, H. Adel, J. Strötgen, Closing the gap: Joint de-identification and concept extraction in the clinical domain, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6945–6952. URL: https://aclanthology.org/2020.acl-main.621. doi:10.18653/v1/2020.acl-main.621.

[17] J. L. Leevy, T. M. Khoshgoftaar, F. Villanustre, Survey on RNN and CRF models for de-identification of medical free text, Journal of Big Data 7 (2020) 73.

[18] I. Pérez-Díez, R. Pérez-Moraga, A. López-Cerdán, J.-M. Salinas-Serrano, M. d. la Iglesia-Vayá, De-identifying spanish medical texts - named entity recognition applied to radiology reports, Journal of Biomedical Semantics 12 (2021) 6.

[19] R. Catelli, V. Casola, G. De Pietro, H. Fujita, M. Esposito, Combining contextualized word representation and sub-document level analysis through bi-lstm+crf architecture for clinical de-identification, Knowledge-Based Systems 213 (2021) 106649. URL: https://www.sciencedirect.com/science/article/pii/S0950705120307784. doi:https://doi.org/10.1016/j.knosys.2020.106649.

[20] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[22] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set, Applied

Soft Computing 97 (2020) 106779. URL: https://www.sciencedirect.com/science/article/pii/S1568494620307171. doi:https://doi.org/10.1016/j.asoc.2020.106779.

[23] R. Catelli, F. Gargiulo, E. Damiano, M. Esposito, G. De Pietro, Clinical de-identification using sub-document analysis and electra, in: 2021 IEEE International Conference on Digital Health (ICDH), 2021, pp. 266–275. doi:10.1109/ICDH52753.2021.00050.

[24] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, A novel covid-19 data set and an effective deep learning approach for the de-identification of italian medical records, IEEE Access 9 (2021) 19097–19110. doi:10.1109/ACCESS.2021.3054479.

[25] K. Murugadoss, A. Rajasekharan, B. Malin, V. Agarwal, S. Bade, J. R. Anderson, J. L. Ross, W. A. Faubion, J. D. Halamka, V. Soundararajan, S. Ardhanari, Building a best-in-class automated de-identification tool for electronic health records through ensemble learning, Patterns 2 (2021) 100255. URL: https://www.sciencedirect.com/science/article/pii/S2666389921000817. doi:https://doi.org/10.1016/j.patter.2021.100255.

[26] C. Meaney, W. Hakimpour, S. Kalia, R. Moineddin, A comparative evaluation of transformer models for de-identification of clinical text data, 2022. URL: https://arxiv.org/abs/2204.07056. doi:10.48550/ARXIV.2204.07056.

[27] A. J. McMurry, B. Fitch, G. Savova, I. S. Kohane, B. Y. Reis, Improved de-identification of physician notes through integrative modeling of both public and private medical text, BMC Medical Informatics and Decision Making 13 (2013) 112.

[28] Z. Jian, X. Guo, S. Liu, H. Ma, S. Zhang, R. Zhang, J. Lei, A cascaded approach for chinese clinical text de-identification with less annotation effort, Journal of Biomedical Informatics 73 (2017) 76–83. URL: https://www.sciencedirect.com/science/article/pii/S1532046417301776. doi:https://doi.org/10.1016/j.jbi.2017.07.017.

[29] Y. Kim, P. Heider, S. Meystre, Ensemble-based methods to improve de-identification of electronic health record narratives, AMIA Annu Symp Proc 2018 (2018) 663–672.

[30] P. Richter-Pechanski, S. Riezler, C. Dieterich, De-Identification of german medical admission notes, Stud Health Technol Inform 253 (2018) 165–169.

[31] Danish Language Council, Retskrivningsordbogen, 4th edition, 2017, including 8 digital issues, Danish Language Council, Bogense, Denmark, 2012.

[32] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 50–61. URL: https://aclanthology.org/2021.naacl-main.5. doi:10.18653/v1/2021.naacl-main.5.

[33] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, H. Xu, CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines, Journal of the American Medical Informatics Association 25 (2017) 331–336. URL: https://doi.org/10.1093/jamia/ocx132. doi:10.1093/jamia/ocx132. arXiv:https://academic.oup.com/jamia/article-pdf/25/3/331/34150625/ocx132.pdf.

[34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational

Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

A. Medical Abbreviations

Table 2
Sources of medical abbreviations

Description	Link	Deep link
List of recognized abbreviations and symbols used in Hospital Sønderjylland	skoletube.dk	https://www.skoletube.dk/download.php?key=MGVhNTVjOTFkZDdlYjlyNDU6ZGI
Symbols and abbreviations used at the laboratory, Hospital Sønderjylland	regionsyddanmark.dk	https://shs.regionsyddanmark.dk/wm272641
Abbreviations and symbols in patient related data	regionsjaellend.dk	http://dok.regionsjaelland.dk/view.aspx?DokID=217238
Definitions and abbreviations, Department of clinical biochemistry, Herlev and Gentofte Hospital	gentofthehospital.dk	https://www.gentofthehospital.dk/afdelinger-og-klinikker/klinisk-biokemisk-afdeling/metodeblade/Documents/Definitioner%20og%20forkortelser%20%E2%80%93%20Klinisk%20Biokemisk%20Afdeling_2020_11_03.pdf
Abbreviations and symbols, psychiatry in Region Nordjylland	pri.rn.dk	https://pri.rn.dk/Assets/15764/Forkortelser-og-symboler-gaeldende-for-sygeplejefaglige-optegnelser.pdf
Medical abbreviations	medicinskefagudtryk.dk	http://medicinskefagudtryk.dk/sidebar/forkortelser/
Ordinary medical abbreviations	medviden.dk	https://www.medviden.dk/vaerktoejer/andre/almindelige-medicinske-forkortelser/
Guidelines for journal writing	gyldendal.dk	http://guga.gyldendal.dk/Munksgaard/sim/media/Munksgaard/Medicinsk%20Fagsprog%202/Retningslinjer.ashx
Abbreviations and designations	medicin.dk	https://pro.medicin.dk/Artikler/Artikel/215
Abbreviations used in Department R, FMB	docplayer.dk	https://docplayer.dk/5008471-Forkortelser-anvendt-i-afd-r-s-instrukser-og-vejledninger-120204-fmh.html
Clinical biochemistry test results	pri.rn.dk	https://pri.rn.dk/Sider/10307.aspx
Family tree abbreviations	pri.rn.dk	https://pri.rn.dk/Sider/29580.aspx
Abbreviations and symbols in the eye speciality	pri.rn.dk	https://pri.rn.dk/Sider/17827.aspx
Accepted abbreviations at vascular surgery department	pri.rn.dk	https://pri.rn.dk/Sider/17293.aspx
Renal abbreviations	pri.rn.dk	https://pri.rn.dk/Sider/16695.aspx
Heart Lung surgery abbreviations	pri.rn.dk	https://pri.rn.dk/Sider/10530.aspx
Intensive Therapy abbreviations	docplayer.dk	https://docplayer.dk/11055021-Hyppigt-anvendte-forkortelser-og-termer-i-intensiv-terapi.html

B. Results Table

Table 3
Distribution of identifiers in each of the training sets automatically annotated with different ratio ceilings and the resulting model and dictionary-based F1 scores.

Ratio ceiling	Sentences with identifiers	Sentences without identifiers	Total sentences	Dictionary-based F1 %	Model F1 %	Name tags	Street tags	Location tags	Total tags
1	21,343	21,343	42,686	33.45	43.86	26,775	1,015	4,309	32,099
2	21,447	21,447	42,894	48.80	65.17	29,003	1,128	3,870	34,001
4	21,131	21,131	42,262	56.02	72.79	28,596	1,217	3,781	33,594
8	21,185	21,185	42,370	61.52	79.19	28,388	1,437	3,749	33,574
16	20,923	20,923	41,846	64.56	82.21	27,988	1,508	3,803	33,299
32	20,597	20,597	41,194	69.11	85.52	27,342	1,660	3,707	32,709
64	19,750	19,750	39,500	71.05	87.36	26,382	2,130	3,533	32,045
128	19,354	19,354	38,708	72.19	87.73	25,963	2,248	3,446	31,657
256	18,058	18,058	36,116	74.02	89.10	24,191	2,391	3,138	29,720
512	16,485	16,485	32,970	74.43	89.62	22,277	2,396	2,897	27,570
1,024	15,490	15,490	30,980	74.42	88.24	21,001	2,323	2,652	25,976
2,048	14,246	14,246	28,492	74.42	88.25	19,617	2,290	2,475	24,382
4,096	13,041	13,041	26,082	73.97	87.85	18,127	2,205	2,332	22,664
8,192	12,424	12,424	24,848	73.82	86.35	17,450	2,196	2,244	21,890
16,384	11,205	11,205	22,410	72.84	87.48	15,826	2,133	1,931	19,890
32,768	10,278	10,278	20,556	72.63	87.06	14,732	2,090	1,756	18,578
65,536	8,879	8,879	17,758	69.77	85.34	12,848	1,915	1,462	16,225
131,072	8,482	8,482	16,964	68.98	86.01	12,330	1,903	1,380	15,613
262,144	8,307	8,307	16,614	68.46	85.37	12,295	1,893	1,291	15,479

C. Ambiguous Performance

Table 4
A comparison of the model and dictionary-based performance on words that occurred both as non-identifiers and identifiers in the test set. Only rows where there is a difference in performance are shown. The total is calculated over all words.

	Non-identifiers		Identifiers	
	Dictionary-based % (total)	**Model % (total)**	**Dictionary-based % (total)**	**Model % (total)**
per	100% (32)	91% (32)	57% (7)	100% (7)
hans	100% (25)	88% (25)	20% (5)	100% (5)
maria	100% (4)	75% (4)	38% (13)	100% (13)
ringe	100% (11)	100% (11)	0% (3)	100% (3)
plads	100% (9)	100% (9)	0% (1)	100% (1)
rask	100% (8)	88% (8)	50% (2)	100% (2)
bak	100% (1)	0% (1)	100% (4)	100% (4)
bo	100% (1)	100% (1)	50% (4)	100% (4)
do	100% (3)	100% (3)	100% (1)	0% (1)
tønder	100% (1)	100% (1)	0% (3)	100% (3)
hammer	100% (1)	0% (1)	100% (2)	100% (2)
rene	100% (1)	100% (1)	50% (2)	100% (2)
slagelse	100% (2)	0% (2)	0% (1)	100% (1)
hammel	100% (1)	100% (1)	0% (1)	100% (1)
land	100% (1)	100% (1)	0% (1)	100% (1)
langeland	100% (1)	0% (1)	0% (1)	100% (1)
stokke	100% (1)	100% (1)	0% (1)	100% (1)
Total	**96% (296)**	**92% (296)**	**50% (88)**	**84% (88)**