Using a Large Open Clinical Corpus for Improved ICD-10 Diagnosis Coding

Anastasios Lamproudis, MSc[1],Therese Olsen Svenning, MSc[1], Torbjørn Torsvik, MD, MPH[1], Taridzo Chomutare, PhD[1,3], Andrius Budrionis, PhD[1,4], Phuong Dinh Ngo, PhD[1,4], Thomas Vakili, MSc[2], Hercules Dalianis, PhD[1,2]
[1]Norwegian Centre for E-health Research, Tromsø, Norway.
[2]Department of Computer and Systems Science (DSV), Stockholm University, Kista, Sweden.
[3]Department of Computer Science, UiT - The Arctic University of Norway, Tromsø, Norway.
[4]Department of Physics and Technology, UiT - The Arctic University of Norway, Tromsø, Norway.

Abstract

With the recent advances in natural language processing and deep learning, the development of tools that can assist medical coders in ICD-10 diagnosis coding and increase their efficiency in coding discharge summaries is significantly more viable than before. To that end, one important component in the development of these models is the datasets used to train them. In this study, such datasets are presented, and it is shown that one of them can be used to develop a BERT-based language model that can consistently perform well in assigning ICD-10 codes to discharge summaries written in Swedish. Most importantly, it can be used in a coding support setup where a tool can recommend potential codes to the coders. This reduces the range of potential codes to consider and, in turn, reduces the workload of the coder. Moreover, the de-identified and pseudonymised dataset is open to use for academic users.

Introduction

One important task in health care when a patient has been diagnosed, treated and discharged from the health care system is to classify the diagnosis for administrative (or financial) and statistical purposes. One standard method used worldwide is the ICD-10 diagnosis coding. ICD stands for *International Statistical Classification of Diseases and Related Health Problems*, and 10 stands for the tenth revision. There are over 30,000 ICD-10 codes divided into 22 Chapters, with each code adhering to a hierarchical structure [1].

The process of selecting codes typically involves trained coders tasked with manually assigning the correct ICD-10 codes to each discharge note. The coding process can sometimes be both tedious and time-consuming. In addition, assigning ICD codes has been shown to be an error-prone process, with over 20-30 percent of codes being erroneous [2]. Therefore, constructing a coding tool to assist the coders would improve both the efficiency and quality of coding. Such a tool can also be used to validate previously assigned codes, which could improve data quality retrospectively.

The aim of this study is to introduce an openly available clinical dataset to train deep learning models for predicting ICD-10 diagnosis codes for discharge summaries written in Swedish in the gastrointestinal surgery domain. Furthermore, it is shown that the predictive performance of the models trained with this dataset is good enough that they can be useful to clinical coders in a clinical setting.

The remainder of this paper is structured as follows. First, the related research is introduced, followed by an introduction of the methods and the data that were used and the inter-annotator agreement levels of the datasets. Next is the results section, followed by the conclusions.

Related Research

Automatic ICD-9 and ICD-10[1] coding has a long history of research, starting with the work by Larkey and Croft [3] in which they combined different classifiers to predict the ICD codes, to the shared task for multi-label ICD-9 classification described by Pestian et al. [4]. The task used a set of 1,954 radiology reports divided into a training

[1]ICD-9 and ICD-10 are different versions of the ICD coding. 9 stands for the ninth version and 10 for the tenth version.

and a test set. One of the best-performing systems in the ICD-9 shared task was developed by Farkas and Szarvas [5]. Their system used a combination of rule-based and machine learning approaches with decision trees and reached an F_1-micro score of 0.89.

Wang et al. [6] employed supervised learning to categorise ICD-10 codes from text documents using natural language processing (NLP) and recurrent neural networks (RNNs). The predictive performance of their ICD-10-CM code classifier reached an F_1-score of 0.62. Also using RNNs, Chen et al. [7] utilised diagnostic data in Chinese as materials and implemented ICD-10 auto-coding using a deep neural network architecture. Previous approaches to predicting gastrointestinal ICD-10 codes for Swedish have demonstrated lower accuracy when predicting full ICD-10 codes, based on previously available datasets [8]. More specifically, an SVM classifier reached an F_1 score of 0.29, outperforming a deep learning classifier based on KB-BERT, a BERT model trained on standard Swedish [9]. Attempting to improve the performance of this classifier, Lövmo [10] used the hierarchical structure of the ICD diagnosis codes but did not find significant improvements.

Data

The data used in this study are comprised of Swedish electronic patient records, where a subset consists of discharge summaries, from several gastrointestinal surgery units from a major Swedish university hospital. Each patient record is assigned one or several ICD-10 diagnosis codes. The records are de-identified to preserve patient privacy. All data were extracted from the Health Bank [11], an infrastructure containing Swedish electronic health records[2][3]. This study uses a baseline dataset (Corpus I) and introduces a new dataset, referred to as Corpus II in the remainder of this paper. Relevant statistics describing the datasets are listed in Table 1, and further details are described below.

Table 1: Characteristics of the two datasets after an initial preprocessing necessary for this work. These statistics might vary slightly in other publications, due to processing steps that might be performed.

	Corpus I	**Corpus II**
Year of release	2021	2022
No. of unique codes	261	415
No. of patients	4,985	113,175
No. of patient records	6,062	317,971
No. of discharge summaries within the dataset	6,062	81,089
No. of tokens	986,000	56,000,000

Baseline Dataset

The first dataset used in this study, henceforth referred to as **Corpus I**, is the **Stockholm EPR Gastro ICD-10 Pseudo Corpus I** introduced by [8]. It consists of 6,062 discharge summaries from 4,985 unique patients, spanning approximately 986,000 tokens. The discharge summaries are distributed amongst 261 unique ICD-10 codes related to gastrointestinal conditions.

Stockholm EPR gastrointestinal ICD-10 Pseudo Corpus II

The dataset introduced in this study is called **Stockholm EPR gastrointestinal ICD-10 Pseudo Corpus II**, and for reasons of simplicity, it will be referred to as **Corpus II** for the rest of this work. **Corpus II** consists of 317,971 patient records, where 81,089 are discharge summaries, for 113,174 patients, spanning nearly 56 million tokens. In total, the records cover 415 unique ICD-10 codes related to gastrointestinal conditions.

[2]Contact the Health Bank, https://www.dsv.su.se/healthbank, at Stockholm University for access to the data. The data is available for academic users.

[3]This research has been approved by the Swedish Ethical Review Authority under permission no. 2022-02386-02 and 2022-02389-02

De-Identification

Data from electronic health records are sensitive and must be handled in a privacy-preserving manner. The corpora used in this study only contain automatically pseudonymised data. The sensitive entities have then been replaced with realistic surrogates, a method that has been shown to preserve the utility of the data for NLP purposes [12, 13]. An example of a de-identified and pseudonymised discharge summary in Swedish can be seen in Figure. 1.

En 82-årig trombylbehandlad man inkommer akut med magsmärtor och ett förmodat lågt Hb. Genomgår 3/3 gastroskopi som visar dels en svårartad esofagit men även ett duodenalulcus. Mår emellertid bra. Ny kontroll av Hb visar cirka 110, mobiliseras, får äta och går hem med recept på trippelbehandling, fortsätter med Omeprazol minst en månad. Inget planerat återbesök.

Translation into English
An 82-year-old man treated with platelets comes in urgently with abdominal pain and a presumed low Hb. Undergoes 3 / 3 gastroscopy, which shows severe esophagitis but also a duodenal ulcer. Feeling well, regardless. New controls of Hb show about 110, mobilised, allowed to eat and goes home with prescription for triple treatment, continuing with Omeprazole for at least a month. No planned return visit.

Manually Assigned ICD-10 Codes
K26.9 Duodenal ulcer, unspecified as acute or chronic, without haemorrhage or perforation. (Main diagnosis).
K21.0 Gastro-esophageal reflux disease with esophagitis. (Side diagnosis).

Figure 1: A de-identified and pseudonymised discharge summary in Swedish also translated to English

Knowledge Representation - Full codes and Block level

A high percentage of classes (ICD-10 codes) in the dataset are represented by very few examples, making the class distribution highly imbalanced. To address the imbalance, multiple labels are condensed into larger classes encompassing multiple codes. These larger classes consist of 10 blocks or categories referred to as **Code Blocks**. The following class scheme was used in the block-level classifiers see Table 2.

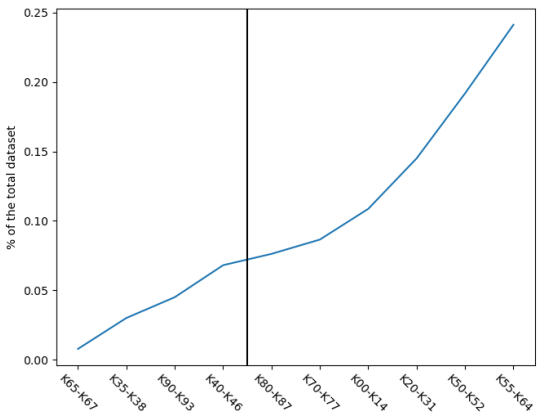
Table 2: ICD-10 Code Blocks of 10 Blocks of Chapter XI

ICD Code Block	Description of Diseases
K00-K14	Diseases of the oral cavity, salivary glands, jaws
K20-K31	Diseases of oesophagus, stomach, duodenum
K35-K38	Diseases of appendix
K40-K46	Hernia
K50-K52	Noninfective enteritis, colitis
K55-K64	Other diseases of intestines
K65-K67	Diseases of peritoneum
K70-K77	Diseases of the liver
K80-K87	Disorders of gallbladder, biliary tract, pancreas
K90-K93	Other diseases of the digestive system

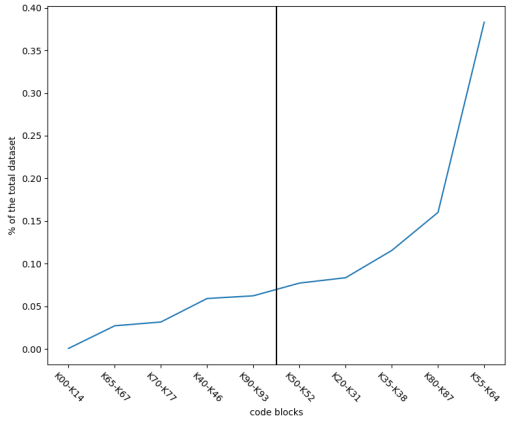
It should be noted that the block-level division is not useful in a practical setting where coders want a specific code assigned to a patient record. However, these code blocks are not random slices of the label space but logical partitions of the gastrointestinal domain, starting from the oral cavity to the rest of the digestive system. Therefore, a block-level classification could be a useful step in building more granular full-code classifiers. We use this block-level division as a benchmark in this paper. In either case, this block-level division is used in this paper as a benchmark and for comparing the results with those obtained by Remmer et al. [8].

Balancing the Corpora

To address the lack of balance in the class distributions within both corpora, the classes that contain too few examples are removed in an effort to balance the datasets. More specifically, the new subsets contain ∼80% of the samples, with the removed 20% corresponding to the samples that belonged to sparsely represented classes. The resulting **Corpus I** 80% subset contains 59 Full Codes, while the resulting **Corpus II** 80% subset contains 69 Full Codes. This means that out of the 261 total codes present in Corpus I and the 415 total codes present in Corpus II, approximately 22% of the codes in Corpus I and 17% in Corpus II are used in 80% of all samples. Figures 2a, 2b, 3a, and 3b present the distribution of the samples to classes along with the cut-off points used to create the balanced versions of the datasets.

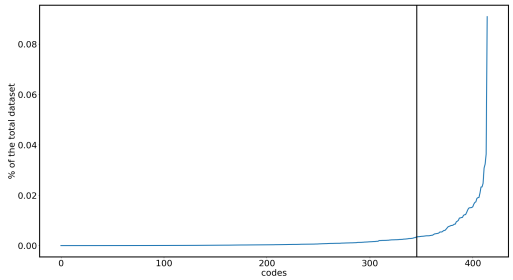


(a) **Corpus II** The blocks to the right of the black line are included in the 80% subset.

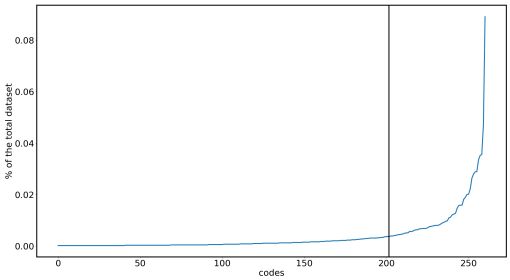


(b) **Corpus I** The blocks to the right of the black line are included in the 80% subset.

Figure 2: The figures illustrate distribution of Code Blocks population for our datasets.



(a) **Corpus II** The blocks to the right of the black line are included in the 80% subset and encompass 69 unique ICD-10 codes.



(b) **Corpus I** The blocks to the right of the black line are included in the 80% subset and encompass 59 unique ICD-10 codes.

Figure 3: The figures illustrate distribution of Full Codes population for our datasets.

Methods

In this study, **Corpus II** is compared with the smaller original counterpart referred to as **Corpus I**. They are both compared in the **Code Blocks** level and in the **Full Codes** level. The model used is SweDeClin-BERT [14], a model pretrained using Swedish clinical text.

SweDeClin-BERT

SweDeClin-BERT [14], which is the model used in this work, is based on KB-BERT [9]. It is adapted to the clinical domain through domain-adaptive continued pretraining, using 17.8 GB of de-identified Swedish clinical text from the Health Bank infrastructure [11]. The model is available to academic researchers.

Training

The training session of SweDeClin-BERT follows the established practices. These practices include extracting 10% of the Corpora as a test set and 90% as a training set. From this training set, a further 10% is isolated as a validation set to use for evaluation during training. The classifier uses the base SweDeClin-BERT model as the core and a classification layer for the final output.

Table 3: Hyper-parameters

Parameter	**Value**
Learning rate	$2 \cdot 10^{-5}$
Batch size	64

As shown in Table 3, a very small learning rate parameter is set, following the example of Devlin et al. [15]. The model is trained using a batch size of 64 until it converges in terms of the validation set loss. The final results are obtained using the held-out test set.

Evaluation

The predictive power of the classifier is evaluated in two different settings. In the first setting, the prediction probability, or confidence of the model, is thresholded at 0.5 and only the predictions with a probability value above this threshold are considered as positive predictions for each class. This approach represents the ”strict” evaluation that makes our results comparable to other studies in terms of model performance metrics.

In the second setting, the five most probable classes for each sample are considered as positive predictions. This is a very pragmatic evaluation approach that was developed in communication with coding experts in Norway and Sweden. ICD coding is not likely to become a fully automated process, and there will always be a need for human evaluation. This evaluation can be thought of as a minimisation of the potential code (label) space to be considered by the human coder. While the coding support is not giving a single correct answer, it can provide the coder with five options that contain the correct code with very high accuracy. The responsibility of picking the correct code from this short list is then left to the coding expert, who can differentiate between similar codes better than the machine does.

A number of different evaluations are performed that can be summarised in the following different setups.

- Train with and evaluate **Corpus I** in the **Full Codes** level.
- Train with and evaluate **Corpus II** in the **Full Codes** level
- Train with and evaluate using the balanced **Corpus I**, comprised of the classes that contain **80%** of the total samples in the **Full Codes** level.
- Train with and evaluate using the balanced **Corpus II**, comprised of the classes that contain **80%** of the total samples in the **Full Codes** level.
- Repeat the previous steps for the **Code Block** level.
- Repeat the previous steps for the **top five** predictions of the model.

Inter-Annotator Agreement Within The Datasets

Inter-annotator agreement (IAA) is the measurement that is used to decide on how difficult a task is for a human or for a machine. The evaluation method is described in [16]. The difficulty of the coding task was investigated by studying annotations by two coding experts from different countries on a random sample of discharge letters from 100 patients. The original coding was carried out by Swedish coders, and the discharge summaries were then re-coded by a senior Norwegian coder. Of the 100 original discharge summaries, 29 were removed because they either did not contain any gastrointestinal information or because they did not contain enough information for re-coding to be possible. Thus, only 71 of the original 100 discharge letters are re-coded.

ICD-10 coding includes many categories, and the K-codes studied here amount to hundreds of available choices for the coders. Another factor that further complicates this task is the possibility of choosing several codes per discharge summary. The chance of different coders making identical categorisation and composition of code sets by chance alone is very small. Correcting for this chance will not have a large impact on the final result, as exemplified in [17].

The coding agreement is assessed using the proportion of agreement. Instead of correcting for a very small chance agreement, we divide the results into three different outcomes for a more detailed assessment; *identical*, *common factor*, and *different*. Identical means that both coders have chosen the same set of codes so that the code sets are identical. Common means that common codes are present, but one coder has some additional codes as well. Different means that there are no common codes in the code set. The levels of agreement for the three different outcomes are listed in Table 4.

Table 4: The percentage agreement between coders analysed on 71 gastrointestinal discharge summaries.

Identical	**Common**	**Different**
0.45	0.27	0.28

Almost half of the code sets in this sample were identical, meaning that the choice and compilation of codes were in full agreement between the coders. If we look at identical and common coding together, the coders are in agreement in 72% of the cases, although they are not identical because one coder included more codes in addition to the common ones. This can be explained by differences in coding rules and practices between Norway and Sweden, especially regarding dependencies between codes. In this study, the coders were not in agreement in 1/3 of the cases. This issue could be improved with the help of a coding assistance tool.

Results

In the first set of experiments, the datasets are compared in the **Full Codes** level. As shown in Table 5, the performance of the classifier varies significantly between using Corpus I and II, and between the different subsets of Corpus II. More specifically, the classifier fails to yield correct predictions for Corpus I. In contrast, the model manages to produce correct predictions for Corpus II with the model trained with the top 80% version yielding significantly better results than the one trained with the full version of the corpus. This is to be expected as removing the least populous classes simplifies the task, both in terms of the number of classes to be predicted along with the remaining classes having enough examples assigned to each of them.

In the second set of experiments, the corpora are compared in the **Code Blocks** level, that is, the 10 grouped categories. In Table 6, the results of the classifier for each of the versions of the two corpora are presented. As these results show, the classifier trained in Corpus II yields better predictions for the complete dataset. On the other hand, the classifier trained on the top 80% version of Corpus I yields better predictions than its counterpart in Corpus II. Again, this is due to the decreased complexity of the task with this modification.

In the next set of experiments, the top five predictive power of the classifiers trained with the different datasets for the

Table 5: The table presents the results for **Full Codes**. The results marked with "-" denote results that are very close to zero and, due to rounding, become zero.

	F_1-score	Precision	Recall
Corpus I	-	-	-
Corpus I (top 80%)	-	-	-
Corpus II	0.61	0.79	0.50
Corpus II (top 80%)	0.71	0.84	0.61

Table 6: The table presents the results for **Code Blocks**

	F_1-score	Precision	Recall
Corpus I	0.87	0.90	0.83
Corpus I (top 80%)	0.94	0.96	0.91
Corpus II	0.89	0.94	0.85
Corpus II (top 80%)	0.92	0.95	0.88

Full Codes is compared and presented in Table 7. In this setup, the classifier trained with Corpus I demonstrates that when considering the 5 most likely predictions, correct predictions are likely to be present, even though they might not have the highest probability according to the model. Regardless, the best classifiers are the ones trained with Corpus II.

Table 7: The table presents the results for **Full Codes** in the top five predictions setup.

	F_1-score	Precision	Recall
Corpus I	0.16	0.16	0.20
Corpus I (top 80%)	0.43	0.44	0.53
Corpus II	0.88	0.88	0.88
Corpus II (top 80%)	**0.94**	0.94	0.94

In the last setup of experimental results presented in Table 8, the classifiers trained in the **code blocks** level, and evaluated in the top five prediction setup, yielded very good performances across all versions of the datasets.

Table 8: The table presents the results for **Code Blocks** in the top five predictions setup.

	F_1-score	Precision	Recall
Corpus I	0.98	0.98	0.98
Corpus I (top 80%)	0.99	0.99	0.99
Corpus II	0.99	0.99	0.99
Corpus II (top 80%)	0.99	0.99	0.99

Discussion

The results for the classifiers differ significantly based on the datasets used for training. The lack of a significant amount of samples in **Corpus I** makes it practically infeasible to develop a classifier for the Full Codes level, limiting the capabilities of every classifier trained with it. This is especially true when evaluated in the "strict" setup, yielding results only when the classifier is trained with the Code Blocks level labels of the dataset. In contrast, using the introduced **Stockholm EPR Gastro ICD-10 Pseudo Corpus II** during training, the classifier can reach adequate performance in both the **Full Codes** level and the **Code Blocks** level.

In contrast, the top-five evaluation shows that the classifier trained with **Corpus I** is able to give a number of correct

predictions within the set. The evaluation also shows that the classifier trained with **Corpus II** (**Stockholm EPR Gastro ICD-10 Pseudo Corpus II**) can reach a very good performance in the top five prediction setup. This classifier has the capability to accurately classify ICD-10 codes in the gastrointestinal domain for Swedish discharge summaries. To further illustrate that, a subset of this openly-available dataset will be used to test this hypothesis in a live demo.

The model's performance for the frequently used codes F_1-score of 0.71, see Table 5, is comparable to the inter-annotator agreement of the human annotators.

Conclusion

In this study, a new de-identified and pseudonymised open dataset of patient records labeled with ICD-10 gastrointestinal codes for the Swedish language is introduced. It is shown that a Swedish BERT-based classifier fine-tuned with this corpus reaches an F_1-score of 0.94 (see Table 7) when choosing among 69 full gastrointestinal ICD codes. This degree of accuracy is high enough that the classifier has the potential for practical use as part of a recommendation tool for coders. It is also shown that even though current state-of-the-art performance on this task is generally unsatisfactory, different presentation methods for the predictions might still yield useful clues for coders in clinical settings.

References

1. WHO. International Classification of Diseases (ICD), https://icd.who.int/browse10/2019/en; 2019. Accessed 2022-12-05. Available from: https://icd.who.int/browse10/2019/en.
2. Jacobsson A, Serdén L. Kodningskvalitet i patientregistret (In Swedish). Socialstyrelsen. 2013. Available from: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/statistik/2013-3-10.pdf.
3. Larkey LS, Croft WB. Combining classifiers in text categorization. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval; 1996. p. 289-97.
4. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A shared task involving multi-label classification of clinical free text. In: Biological, translational, and clinical language processing. Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 97-104. Available from: https://www.aclweb.org/anthology/W07-1013.
5. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics. 2008 4;9(3):S10. Available from: https://doi.org/10.1186/1471-2105-9-S3-S10.
6. Wang Sm, Chang Yh, Kuo Lc, Lai F, Chen Yn, Yu Fy, et al. Using deep learning for automatic ICD-10 classification from free-text data. European Journal of Biomedical Informatics. 2020;16(1).
7. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning [Journal Article]. JMIR Med Inform. 2021;9(8):e23230. Available from: https://medinform.jmir.org/2021/8/e23230https://doi.org/10.2196/23230http://www.ncbi.nlm.nih.gov/pubmed/34463639.
8. Remmer S, Lamproudis A, Dalianis H. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In: Proceedings of RANLP 2021: Recent Advances in Natural Language Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria; 2021. p. 1158-66.
9. Malmsten M, Börjeson L, Haffenden C. Playing with Words at the National Library of Sweden–Making a Swedish BERT. arXiv preprint arXiv:200701658. 2020.
10. Lövmo E. Hierarchical Diagnosis Code Classification of Discharge Letters using Distributional Semantics. Stockholm University,https://daisy.dsv.su.se/fil/visa?id=241830; 2022.
11. Dalianis H, Henriksson A, Kvist M, Velupillai S, Weegar R. HEALTH BANK- A Workbench for Data Science Applications in Healthcare. Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015), J Krogstie, G Juell-Skielse and V Kabilan, (Eds), Stockholm, Sweden, June 11, 2015, CEUR Workshop Proceedings. 2015:34-44. Available from: http://ceur-ws.org/Vol-1381/paper1.pdf.
12. Berg H, Henriksson A, Dalianis H. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Louhi 2020, in conjunction with EMNLP 2020; 2020. p. 1-11.
13. Vakili T, Dalianis H. Utility Preservation of Clinical Text After De-Identification. In: Proceedings of the 21st

Workshop on Biomedical Language Processing at ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 383-8. Available from: https://aclanthology.org/2022.bionlp-1.38.
14. Vakili T, Lamproudis A, Henriksson A, Dalianis H. Downstream task performance of bert models pre-trained using automatically de-identified clinical data. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 2022. p. 4245-52.
15. Devlin J, Chang MW, Lee K, Toutanova K. 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); 2019. p. 4171-86.
16. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Computational linguistics. 2008;34(4):555-96.
17. Peng M, Eastwood C, Boxill A, Jolley RJ, Rutherford L, Carlson K, et al. Coding reliability and agreement of International Classification of Disease, 10th revision (ICD-10) codes in emergency department data. International Journal of Population Data Science. 2018;3(1).