Abstract

Clinical Text Notes (CTNs) contain physicians’ reasoning process, written in an unstructured free text format, as they examine and interview patients. In recent years, several studies have been published that provide evidence for the utility of machine learning for predicting doctors’ diagnoses from CTNs, a task known as ICD coding. Data annotation is time consuming, particularly when a degree of specialization is needed, as is the case for medical data. This paper presents a method of augmenting a sparsely annotated dataset of Icelandic CTNs with a machine-learned data imputation in a semi-self-supervised manner. We train a neural network on a small set of annotated CTNs and use it to extract clinical features from a set of un-annotated CTNs. These clinical features consist of answers to about a thousand potential questions that a physician might find the answers to during a consultation with a patient. The features are then used to train a classifier for the diagnosis of certain types of diseases. We report the results of an evaluation of this data augmentation method over three tiers of information that are available to a physician. Our data augmentation method shows a significant positive effect, which is diminished when an increasing number of clinical features, from the examination of the patient and diagnostics, are made available. We recommend our method for augmenting scarce datasets for systems that take decisions based on clinical features that do not include examinations or tests.

1 Introduction

When a patient consults a physician, communication is created in the patient’s medical records. The physician notes down the patient’s signs, symptoms, results of physical examination, the clinical thinking process, and if any diagnostic tests are warranted – in a free text format known as a Clinical Text Note (CTN). Then, the physician saves the diagnoses, using the International Classification of Diseases (ICD) code, that they made during the consultation. Thus, each CTN contains free text, from which clinical features can be extracted, in addition to the ICD classification code.

Previous work has shown the benefits of training machine learning classifiers on clinical features for automated ICD coding (Liang et al., 2019; Ellertsson et al., 2021; Zhang et al., 2020; Pascual et al., 2021; Kaur et al., 2021; Blanco et al., 2021). Ellertsson et al. (2021) hand-annotated features in 800 CTNs and trained a classifier to predict ICD codes for one of four types of primary headache diagnoses. Liang et al. (2019) hand-annotated a significantly larger set, i.e. about 6,000 CTNs, for the purpose of training a classifier to predict various types of diseases, i.e. 55 ICD codes in total. Additionally, Liang et al. developed a clinical feature extraction model (CFEM), for the purpose of automatically extracting features from the CTNs.

On its own, the CFEM is beneficial because it could solve the common clinical problem of getting a quick and comprehensive overview of a patient, when meeting a clinician for the first time. A clinician could search a patient’s medical history with a question such as “Has the patient ever had a colonoscopy?”. The ICD classifiers have, on the other hand, the potential of being integrated into a Clinical Decision Support System (CDSS), where they could, for example, predict if a physician should order an MRI for a patient when presented with a particular symptom, what kind of blood tests are warranted, or any other diagnostic test for that matter.

Generally, machine learning systems require large quantities of training data (Hlynsson et al., 2019) and ICD classifiers are no exception. In order to develop a high accuracy ICD classifier, without annotating large amount of CTNs, we experiment with a method of: 1) annotating a small subset of the CTNs with question-answer pairs which are used for training the CFEM, and then 2) use the
trained feature extractor to extract clinical features from a larger dataset of CTNs for training the classifier to predict one out of six ICD codes\(^1\). We call this method semi-self-supervised because it lies at the intersection of 1) semi-supervised learning, which combines a small amount of labeled data with large amounts of unlabeled data (Van Engelen and Hoos, 2020) and 2) self-supervised learning, which learns to predict missing parts of inputs (Mao, 2020).

In prior work on ICD coding, classifiers are trained on discharge summaries, after the patient has left the clinic (Liang et al., 2019; Zhang et al., 2020; Pascual et al., 2021; Kaur et al., 2021; Blanco et al., 2021). We instead focus on evaluating our model on stages in the primary health care pipeline where the recommendations of machine learning models would be the most effective. We thus introduce a novel three-tiered evaluation system that is designed to mirror the circumstances where ICD classification methods would actually be used and we evaluate our semi-self-supervised data augmentation method on these three tiers: 1) before the patient meets a physician, 2) after the physician performs the patient examination, and 3) after the physician has ordered diagnostic tests.

Our evaluation results show that the data augmentation method has a significant benefit for tier 1, i.e. before the patient meets a physician, but not for the other two.

2 Related Work

Liang et al. (2019) frame the problem of clinical feature extraction from CTNs as a question-answering task. Every clinical feature mentioned in a given CTN is marked, as well as the start and end of the text span referring to a given clinical feature. A question is saved in the context of the text span, which contains the answer to that specific question. For example, given the text span “the patient has a fever”, the question “Does the patient have a fever?” is saved with a binary value of 1. Out of 1.3 million CTNs from a single institution in China, Liang et al. annotated about 6,000 CTNs for training a CFEM, based on a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) enriched with word embeddings. The feature extractor is trained on a batch of (CTN, question, text span) tuples as input with the goal of optimizing for the text span that contains the corresponding answer to the question in the given CTN. Thereby, the model learns to extract relevant clinical features from the questions put forward in the context of the CTN. Liang et al. used the CFEM to extract features from the whole set of un-annotated CTNs. The extracted features were then used to train a classifier, based on multi-class logistic regression, to predict an ICD code from a set of 55 codes.

Ellertsson et al. (2021) hand-annotated clinical features (in a similar manner as Liang et al.) in 800 CTNs from a common medical database of all primary care clinics in Iceland. Each CTN had an accompanying ICD code for one of four types of headache diagnoses. The resulting features (text spans) were then used to train a Random Forest classifier, for predicting one of the four possible ICD codes. Furthermore, they performed a retrospective study where the classifier was shown to outperform general practitioners on the four types of headache diagnostics.

In this paper, we expand upon the work of Ellertsson et al. The main difference between our work and theirs can be summarized as follows:

- We do not compare our ICD classifier to general practitioners.
- We hand-annotate questions-answers pairs in 2,422 CTNs, which includes a larger number of ICD codes, 42 in total (see Table 4 in the Appendix).
- Using the hand-annotated CTNs, we train CFEMs, based on Transformer models (Vaswani et al., 2017), for extracting clinical features, and compare them to a couple of LSTM models. These feature extractors are used to extract features from un-annotated CTNS as well as annotated CTNs.
- We perform a three-tiered evaluation of our classifiers on six of the ICD codes for pediatric (under 18) patients (see Table 5 in the Appendix).

Transformer-based models have rapidly become a popular choice for automated ICD coding. These models have been trained on CTNs in a fully end-to-end manner (Zhang et al., 2020; Pascual et al., 2021; Kaur et al., 2021; Blanco et al., 2021). A
drawback of this approach is that physicians will often write down their hypothesized diagnoses which injects a serious bias to the data, a problem that our approach, of using one model for clinical feature extraction and another for clinical prediction, circumvents. For example, a fully end-to-end machine learning model might learn to associate the qualitative comment by a physician “the patient probably has a migraine without aura” in a patient with a migraine-without-aura ICD code. Our method avoids this by creating a bottleneck of information, where only specific questions are being answered.

Our approach also opens the door for interpreting the results of the ICD classifier, as the importance of each input feature to the classifier can be visualized, for example by portraying input coefficients in the case of linear models (e.g. logistic regression) or plotting other interpretability metrics, such as SHAP values (Lundberg and Lee, 2017).

### 3 Approach

#### 3.1 Data and annotation

We use the dataset from the same source as Ellertsson et al. (2021), i.e. from the Primary Health Care Service of the Capital Area (PHCCA) in Iceland. The dataset consists of 1.2 million CTNs, written in Icelandic, from 200 thousand unique patients that were collected in clinical consultations taking place from January 2006 to April 2020. Physicians are instructed not to write anything that can uniquely identify their patients in the notes, but we also used a combination of a parsing system for Icelandic (Porseinsson et al., 2019) as well as a regex command to remove any personally identifiable information, such as names, personal identification numbers and phone numbers. This dataset contains CTNs that have an associated ICD code, but consist otherwise of unstructured text from which clinical features can be extracted.

In the same manner as described by Ellertsson et al., we reduced the full dataset by applying a filter which only keeps notes that contain any word from a medical keyword dictionary. From this reduced dataset, we randomly selected 2,422 notes which were manually annotated by a physician, resulting in question-answer pairs as described in Section 2.

As an example annotation, for a CTN containing the text “the patient is not coughing”, one clinical feature is the pair consisting of the question “does the patient have a cough?” and the binary-valued answer “0”, with the corresponding text span “not coughing”. Some answers are continuous-valued, such as for the question “what is the patient’s blood pressure?”.

The number of clinical features that we use to train the extraction model to output is 942. There is typically a heavy class imbalance for each feature, where the binary questions have on average a 0.75 positive answer ratio, with a standard deviation of 0.2. The reason for this sweeping class imbalance is that physicians generally only ask questions that are relevant and with an affirmative answer.

For our three-tiered classifier evaluation, we define three strict subsets of these features, as described in Section 3.6. Each question is also paired with another binary variable which indicates whether an answer to that question can be found in the CTN or not.

The dataset is split into adults, that are 18 years old or older, and children. Within each age group, 80% of the dataset is allocated for training, 10% for development/validation, and hold out 10% for final testing (see Table 1). The split is stratified to ensure that each set has an equal proportion of sexes and ICD codes.

---

2The annotator is a white Icelandic male physician in his thirties, specializing in general practice / family medicine.
3.2 Pre-trained Transformer-based models

We compared four existing Transformer-based models in our experiments, based on the ELECTRA (Clark et al., 2020) and RoBERTa (Liu et al., 2019) architectures. We evaluated an ELECTRA-small\textsuperscript{3}, ELECTRA-base\textsuperscript{4} and two RoBERTa-base models\textsuperscript{5,6} (consisting of 14M, 110M and 125M parameters, respectively). All models have been pre-trained on the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018), which consists of approximately 1.69B tokens from genres such as news articles, parliamentary speeches, novels and blogs. For one of the RoBERTa models, which we refer to as RoBERTa+, the IGC was supplemented with texts obtained from online sources, increasing the size of the pre-training corpus to 2.7B tokens.

The RoBERTa models were pre-trained for 225k steps with a batch size of 2k. Otherwise, all models were pre-trained using default settings. The pre-training process and additional training data for the RoBERTa models is described in further detail by Snæbjarnarson et al. (2022).

3.3 LSTM architectures

For a baseline comparison, we created two LSTM models. The first one (LSTM 1) tokenizes and trains the embeddings from scratch, whereas the second one (LSTM 2) pre-processes the inputs with GloVe (Pennington et al., 2014) embeddings.

3.3.1 LSTM 1

The model splits up the tokenized input into question and content parts. The content part gets a 256-dimensional embedding and the question gets a 32-dimensional embedding. Each embedding is then passed to its own, uniquely parameterized two-layer bi-directional LSTM model, where each layer has 256 units.

The outputs from those two parts are then concatenated and used to 1) train a set of dense networks, where one is tasked with predicting whether an answer to the question can be found in the text and, if yes, the other dense network predicts the probability of the answer being affirmative (in the case of binary questions), and 2) predict the start and end indices of the tokens that mark the span of the answer in the context part.

3.3.2 LSTM 2

LSTM 2 has the same architecture as LSTM 1, except there is no embedding layer and the inputs have been processed by a pre-trained GloVe model. The GloVe embeddings\textsuperscript{7} where pre-trained on the IGC.

3.4 Clinical feature extraction models

We fine-tuned the four Transformer-based models, mentioned in Section 3.2, on the hand-annotated data in order to develop a CFEM. The fine-tuning was carried out in the following manner: starting with the pre-trained transformers weights, the top layer was replaced with a randomly initialized network, and the whole system was then trained end-to-end for question-answering. We also trained the two LSTM models described in Section 3.3 from scratch for a CFEM comparison.

Each model learns to output the answer span for each question\textsuperscript{8} as well as the probability of the answer being affirmative for binary-valued questions. The Transformer-based models were defined and trained using the Transformers (Wolf et al., 2019) and PyTorch libraries (Paszke et al., 2019) and the LSTM models were defined and trained using TensorFlow (Abadi et al., 2016).

3.5 Semi-self-supervised learning

Once our CFEMs were trained, we saved their outputs over all the CTNs (i.e. 2,422 annotated CTNs used for training and 750 randomly selected un-annotated CTNs) to disk. The outputs define the matrix of independent variables $X$ which is, along with the dependent variable array $y$ of ICD codes, used to train our logistic regression ICD classifier (implemented in scikit-learn (Pedregosa et al., 2011)).

CTNs require expertise to interpret, which results in a high cost when labelling medical datasets. This is especially true for AI researchers that are working with a language with much fewer resources than English (Blanco et al., 2021), such as Icelandic.

\textsuperscript{3}https://huggingface.co/jonfd/electra-small-igc-is. CC-BY-4.0 license.
\textsuperscript{4}https://huggingface.co/jonfd/electra-base-igc-is. CC-BY-4.0 license.
\textsuperscript{5}https://huggingface.co/mideind/IceBERT. AGPL 3.0 license.
\textsuperscript{6}https://huggingface.co/mideind/IceBERT-igc. AGPL 3.0 license.
\textsuperscript{7}https://github.com/stofnun-arna-magnussonar/ordgreypingar_embeddings/tree/main/GloVe
\textsuperscript{8}If the question is not answered in the CTN, the model outputs an impossible span in the text, which is technically implemented as starting at the $0^{th}$ token (a special “start” token) and the $1^{st}$ token of the proper context.
Figure 1: Leveraging a Sparsely Annotated Dataset. Our clinical feature extraction model learns to mark text spans (clinical features), containing an answer to a set of given clinical questions, from CTNs in which answer spans have been hand-annotated. The feature extractor is then used to extract answer spans – given the same set of questions – from a large set of CTNs that have diagnoses (ICD codes), but no marked answer spans. Finally, the extracted answer spans are used to train the ICD classifier. In this way, we make full use of a large set of CTNs that is only partly annotated and combine it with a much smaller set of human-annotated CTNs to learn automated ICD coding.

In our project, we have a large collection of CTNs, each of which is marked with a doctor’s diagnosis, but does not contain answer spans for the set of questions for our clinical features. We input the un-annotated CTNs to a CFEM, that is trained on a much smaller subset of the data, to take advantage of the supervisory signal offered by the ICD code of each un-annotated CTN. This step keeps the interpretable clinical features and removes potential bias from the CTNs. This set of CTNs with imputed clinical feature values is then combined with our “gold standard” set of annotated CTNs, and both are used for training the ICD classifier (see Figure 1).

3.6 Three-tiered evaluation
To simulate the different stages of a physician’s evaluation of a patient in real clinical circumstances, we limit the number of features that are available to the classifier at each stage:

- **Tier 1:** Before a patient meets with a physician. This includes the patient’s main complaint, history, symptoms, and vital signs (420 features).
- **Tier 2:** After the patient has been examined by a physician (582 features).
- **Tier 3:** After results from diagnostics are available (608 features).

The full list of features is provided in the Appendix: Table 6 and Table 7 for tier 1, which are features that the patient could self-report. Tables 8 and 9 show the features for tiers 2 and 3, respectively. After tiers 2 and 3, decisions need to be taken regarding what further tests need to be ordered, for example imaging.

Note that our system could fit into a triage context at tier 1. The patient could fill out an online questionnaire and get recommendations depending on the results, for example, to go to the emergency room, to go the general physician, or maybe just rest at home with a set of self-care instructions.

4 Results and Discussion
4.1 Clinical feature extraction model training
The CFEMs were trained over three epochs on the subset of hand-annotated CTNs (see Table 1). For the ELECTRA-base and RoBERTa-base transformers, each epoch took approximately eight hours on Cloud TPU v3 with eight cores, and half that for...
ELECTRA-small. The training took approximately three hours for each epoch for the LSTMs.

The RoBERTa+ model, which is pre-trained on the largest corpus, achieves the best results for all three metrics that we monitor (see Table 2): a span-based $F_1$-score, to evaluate the question-answering portion of the models, and the Matthews correlation coefficient (MCC) (Matthews, 1975; Chicco and Jurman, 2020) for the binary-valued clinical features (Binary MCC) and for predicting whether the question is answered in the text (Impossible MCC).

We chose the MCC metric because it is appropriate for imbalanced data (Chicco, 2017) (see discussion of our data in Section 3.1) and it offers a suitable combination of the four confusion matrix metrics: true positives, true negatives, false positives and false negatives.

Note in Table 2 that the high $F_1$-scores are due to the fact that most questions were correctly predicted to be not answered in any given context. This could be due to the fact that the 15.8 GB corpus, which was used to train RoBERTa+, contains 33 MBs of medical texts. Although this is not a large proportion, it could be enough for the model to have learned transferable representations of medical vocabulary.

To our surprise, the ELECTRA-base model was outperformed by RoBERTa (both are trained on equal-sized corpora), even though ELECTRA has, previously, been shown to outperform RoBERTa on question-answering tasks (Clark et al., 2020).

The LSTM variation whose inputs were not pre-processed by a pre-trained GloVe model (LSTM 1) performed better according to the MCC metrics (but slightly worse according to the $F_1$-score) than the other (LSTM 2). We hypothesize that it is due to the fact that the pre-trained embeddings are not trained with any tokenization, but rather on whole words. The free-text style of doctor’s notes can include words or abbreviations that are not defined for the GloVe embeddings.

### 4.2 ICD classifier training

#### 4.2.1 Transformer vs. LSTM

After training and evaluating the CFEMs, we validated the data augmentation scheme described in Section 3.5. We used the best-performing models from each category, RoBERTa+ and LSTM 1, to extract the clinical features from the children’s notes. These features, along with their associated ICD codes, were then used to train the classifier.

Table 3 shows the diagnostic metrics of the classifier for tier 3 depending on the feature extractor. Using RoBERTa+ yielded a higher weighted average for all diagnostic metrics compared to LSTM 1.

#### 4.2.2 Qualitative analysis

To verify that the relationship between our features and the outputs of our models matches our clinical intuition, we use SHAP (Shapley additive explanation) values (Shapley, 1953) to show the impact of each feature in the prediction of our logistic regression classifier, trained on the features in tier 3 extracted by RoBERTa+.

The feature importance plot is shown in Figure 2. We see, for example, that the top four features are headache-related features and contribute to classifying a CTN as Tension-type headache, migraine with- and without aura. The two top features after that involve the doctor doing a physical examination of the patient’s lung and contribute to predicting whether the patient has pneumonia or bronchitis. The sixth most impactful feature is then the result of an examination of the patient’s ear, the result of which contributes to the diagnosis of Otitis media (a disease of the middle ear).

#### 4.2.3 Data augmentation experiment

In the next set of experiments, we investigated the effect of augmenting a data set consisting of 303 human-labeled children’s CTNs with a varying number of machine-labeled children’s CTNs for the purpose of training an ICD classifier.

We trained logistic regression classifiers using 5-fold cross-validation over the whole children set. Each classifier had L1 regularization with the in-

---

$^9$ Due to time constraints, our evaluation of the data augmentation method is limited to only using the children CTNs.
Table 3: Detailed ICD classification metrics. Per-class metrics for clinical diagnosis prediction when a logistic regression classifier is trained on features extracted from CTNs by either our RoBERTa+ transformer or the baseline LSTM 1 model. MCC is the Matthews correlation coefficient, TPR is the true positive rate and TNR is the true negative rate.

| Condition                  | RoBERTa+   | LSTM 1     |
|----------------------------|------------|------------|
|                            | $F_1$-score| MCC  | TPR  | TNR  | $F_1$-score| MCC  | TPR  | TNR  |
| Migraine without aura      | 0.40       | 0.36      | 0.33  | 0.97  | 0.00       | 0.00  | 0.00  | 1.00  |
| Migraine with aura         | 0.67       | 0.70      | 0.50  | 1.00  | 0.40       | 0.36  | 0.33  | 0.97  |
| Tension-type headache      | 0.94       | 0.89      | 1.00  | 0.88  | 0.86       | 0.73  | 1.00  | 0.71  |
| Otitis media, unspecified  | 0.00       | 0.00      | 0.00  | 1.00  | 0.57       | 0.60  | 1.00  | 0.90  |
| Bacterial pneumonia        | 0.86       | 0.83      | 1.00  | 0.93  | 0.75       | 0.75  | 0.60  | 1.00  |
| Acute bronchitis           | 1.00       | 1.00      | 1.00  | 1.00  | 0.33       | 0.29  | 0.25  | 0.97  |
| Weighted average           | **0.81**   | **0.78**  | **0.85** | **0.85** | **0.64**  | **0.56**  | **0.70**  | **0.70** |

Figure 2: Feature importance plot. The features are scored by their SHAP values. The size of the colored bar in each feature’s row indicates the contribution of that feature to predicting the disease with the corresponding color.

Conclusions and Future Work

Our results show that training a CFEM on a small annotated subset of CTNs and use it to extract features from a larger, un-annotated dataset can increase the performance of an ICD classifier. However, the effect is only positive and significant in the context before a patient has been examined by the physician.

A future line of work is to further validate different classifiers by performing prospective studies which allow us to get insight into how the classifier performs in real clinical situations. This can be done by integrating the classifier into a CDSS, where a patient can log into a secure portal, at home or at a medical institution, and answer targeted questions regarding their symptoms. The CDSS could build a list of differential diagnoses, recommend further diagnostics based on the patients symptoms, and then write out the CTN for the clinician. This does not disturb the clinical workflow, saves time for medical staff and poten-
Figure 3: Data Augmentation Results. Each classifier is trained on fixed set of hand-annotated clinical features, in addition to a varying number of features automatically extracted by the RoBERTa+ model, i.e. machine-labeled features. There are 237 hand-annotated CTNs in each training set and each step along the x-axis adds 75 machine-labeled CTNs. Each point in the augmented curves shows the cross-validated metrics (accuracy in the left column and MCC in the right column) averaged over 20 random subsets of machine-labeled points that are added to the training set and the error band (the colored area around the Augmented Roberta+) signifies the 95% confidence intervals. The dashed lines indicate the performance of the classifiers trained only on hand-annotated data.

Acknowledgements

Acknowledgements will appear in the final version.
References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16), pages 265–283.

Alberto Blanco, Sonja Remmer, Alicia Pérez, Hercules Dalianis, and Arantza Casillas. 2021. On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages with Fewer Resources than English. In RANLP 2021: Recent Advances in Natural Language Processing, 1-3 Sept 2021, Varna, Bulgaria, pages 165–172. Association for Computational Linguistics.

Davide Chicco. 2017. Ten quick tips for machine learning in computational biology. BioData mining, 10(1):1–17.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1):1–13.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv preprint arXiv:2003.10555.

Steindor Ellertsson, Hrafn Loftsson, and Emil L. Sigurdsson. 2021. Artificial intelligence in the GP office: a retrospective study on diagnostic accuracy. Scandinavian Journal of Primary Health Care, 39(4):448–458.

Hlynur Hlynsson, Alberto Escalante-B., and Laurenz Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Grammar for Icelandic and an Accompanying Parsing System. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 1397–1404.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 32:8026–8037.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30.

Scott M Lundberg and Su-In Lee. 2021. A Survey on Self-supervised Pre-training for Sequential Transfer Learning in Neural Networks. arXiv preprint arXiv:2007.00800.

Huanru Henry Mao. 2020. A Survey on Self-supervised Learning: Limitations and Opportunities. arXiv preprint arXiv:2104.06709.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 32:8026–8037.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Lloyd S Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, Contributions to the Theory of Games II, pages 307–317. Princeton University Press, Princeton.

Vésteinn Snaebjarnarson, Haukur Barri Simonarson, Pétur Örri Ragnarsson, Svanhvit Lílja Ingólfsdóttir, Haukur Páll Jónsson, Vilhjálmur Porsteinsson, and Hafsteinn Einarrsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In Proceedings of the 13th International Conference on Natural Language Resources and Evaluation (LREC 2022), Marseille, France.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic
A Appendix

| ICD code | Description |
|----------|-------------|
| G43.0    | Migraine without aura |
| G43.1    | Migraine with aura |
| G44.0    | Cluster headaches and other trigeminal autonomic cephalgias |
| G44.2    | Tension-type headache |
| G44.4    | Drug-induced headache, not elsewhere classified |
| G45.9    | Transient cerebral ischemic attack, unspecified |
| H66.0    | Acute suppurative otitis media |
| H66.9    | Otitis media, unspecified |
| I10      | Essential (Primary) Hypertension |
| I63.0a   | Cerebral infarction |
| I63.1    | Cerebral infarction |
| I63.2a   | Cerebral infarction due to unsp. occl. or stenosis of precerebral arts. |
| I63.3    | Cerebral infarction due to thrombosis of cerebral arts. |
| I63.4    | Cerebral infarction due to embolism of cerebral arteries. |
| I63.5    | Cerebral infarction due to unsp. occl. or stenosis of cerebral arts. |
| I63.6    | Cerebral infarction due to cerebral venous thrombosis, nonpyogenic |
| I63.8    | Other cerebral infarction |
| I63.9    | Cerebral infarction, unspecified |
| I84.0    | Haemorrhoids |
| J00      | Acute nasopharyngitis [common cold] |
| J01      | Acute sinusitis |
| J01.0    | Acute maxillary sinusitis |
| J01.9    | Acute sinusitis |
| J02.0    | Streptococcal pharyngitis |
| J03.0    | Streptococcal tonsillitis |
| J03.9    | Acute tonsillitis |
| J05.0    | Acute obstructive laryngitis |
| J10.1    | Influenza due to other identified influenza virus w/ other resp. manifs. |
| J11.1    | Influenza with other resp. manifs., virus not identified |
| J12.9    | Viral pneumonia, unspecified |
| J15.9    | Bacterial pneumonia, unspecified |
| J15.0    | Bacterial pneumonia, not elsewhere classified |
| J15.1    | Pneumonia due to Mycoplasma pneumoniae |
| J15.8    | Pneumonia due to other specified bacteria |
| J15.9    | Bacterial pneumonia, unspecified |
| J20.9    | Acute bronchitis |
| J44.1    | Chronic obstructive pulmonary disease with (acute) exacerbation |
| J44.9    | Chronic obstructive pulmonary disease, unspecified |
| J45.0    | Predominantly allergic asthma |
| J45.9    | Asthma, unspecified |
| M54.1a   | Radiculopathy |
| M54.5a   | Low back pain |
| S83.2    | Tear of meniscus, current injury |

Table 4: ICD codes associated with notes used during training of the clinical feature extraction model.

Table 5: ICD codes associated with notes used during classifier training.
| History of migraines | History smoking | History smoking package years | History of wiplash | History of alcoholism | History is regularly active |
|---------------------|----------------|-------------------------------|------------------|----------------------|---------------------------|
| History of head trauma | History of hypertension | History smoking active time | History smoking packages per day | History known allergy | History of bells palsy |
| History of allergic rhinitis | History of hyperthyroidism | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of hypothyrosis |
| History of abdominal pain | History of headaches | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of palpitations |
| History of chest pain | History of the natural course of the disease | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of chronic sinusitis |
| History of cough | History of ear pain | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of dermatitis | History of fever | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of emphysema | History of flu | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of fever | History of gout | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of gout | History of headache | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of headache | History of hypertension | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of hypothyroidism | History of insomnia | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of anxiety | History of migraines | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of multiple sclerosis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of diabetes mellitus | History of nystagmus | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of hypothyroidism | History of otitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of wiplash | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of alcoholism | History of puffiness | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of allergy | History of quiescence | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of asthma | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of allergic rhinitis | History of tinnitus | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of allergic rhinitis | History of vomiting | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of allergy | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of wiplash | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |
| History of sinusitis | History of sinusitis | History smoking active time | History smoking packages per day | History of recurrent sinusitis | History of depression |

Table 7: Tier 1 features, Part 2 of 2.
Table 8: Tier 2 features. This tier also includes the previous tier’s features.

| Examination lung auscultation a | Examination proprioception abnormal | Examination is obese | Examination palpable neck lymph | Examination heart auscultation | Examination systolic heart murm | Examination lung auscultation b | Examination lung auscultation c | Examination lung auscultation d |
|---------------------------------|-------------------------------------|----------------------|----------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| Examination abnormal or absent  | Examination abnormal neurologic     | Examination abnormal or absent | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination abdominal gait      | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination prolapse-abnormal   | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination neck stiffness      | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination generally sick      | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination is walking abd      | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination shoulder muscles pa| Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination abnormal or reduced | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation e | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation f | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation g | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation h | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation i | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation j | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation k | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation l | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation m | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation n | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation o | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation p | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation q | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation r | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation s | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation t | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation u | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation v | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation w | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation x | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation y | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |
| Examination lung auscultation z | Examination normal                | Examination normal     | Examination systolic heart murm | Examination systolic heart murm | Examination lung auscultation c | Examination abnormal or absent  | Examination abnormal or absent  | Examination abnormal or absent  |

Table 9: Tier 3 features. This tier also includes the two previous tiers’ features.