Self-Supervised Detection of Contextual Synonyms in a Multi-Class Setting: Phenotype Annotation Use Case

Jingqing Zhang[1,2], **Luis Bolanos**[2], **Tong Li**[2], **Ashwani Tanwar**[2], **Guilherme Freire**[2]
Xian Yang[3], **Julia Ive**[1], **Vibhor Gupta**[2], **Yike Guo**[1,2,3]
[1]Data Science Institute, Imperial College London, UK
[2]Pangaea Data Limited, UK, USA
[3]Hong Kong Baptist University, Hong Kong SAR, China
jzhang,lbolanos,tli,atanwar,gfreire,vgupta@pangaeadata.ai
j.ive,y.guo@imperial.ac.uk xianyang@comp.hkbu.edu.hk

Abstract

Contextualised word embeddings is a powerful tool to detect contextual synonyms. However, most of the current state-of-the-art (SOTA) deep learning concept extraction methods remain supervised and underexploit the potential of the context. In this paper, we propose a self-supervised pre-training approach which is able to detect contextual synonyms of concepts being training on the data created by shallow matching. We apply our methodology in the sparse multi-class setting (over 15,000 concepts) to extract phenotype information from electronic health records. We further investigate data augmentation techniques to address the problem of the class sparsity. Our approach achieves a new SOTA for the unsupervised phenotype concept annotation on clinical text on F1 and Recall outperforming the previous SOTA with a gain of up to 4.5 and 4.0 absolute points, respectively. After fine-tuning with as little as 20% of the labelled data, we also outperform BioBERT and ClinicalBERT. The extrinsic evaluation on three ICU benchmarks also shows the benefit of using the phenotypes annotated by our model as features.

1 Introduction

Supervised fine-tuning on the top of the BERT-based models has recently become the standard approach in Information Extraction delivering state-of-the-art (SOTA) results across different tasks (Devlin et al., 2019). The dependence of these models on the availability of the costly human-annotations remains a serious obstacle towards a large scale deployment of such models. This problem is especially actual in the clinical domain with limited availability of experts.

In the self-supervised setting, automatic annotations are cheap to produce. Some rule-based automatic labelers, such as CheXpert (Irvin et al., 2019), which is built on NegBio (Peng et al., 2018), are often used to create training data for supervised BERT-based models (e.g., automatic annotators of radiology reports (Smit et al., 2020)). Those models usually generalise over a small set of classes (under 20).

Other automatic labelers exploit ontologies. For example, the UMLS (Unified Medical Language System) ontology (Bodenreider, 2004b) is almost predominantly used to match linguistics patterns in clinical text to medical concepts (e.g., using the MetaMap tool (Aronson, 2006)). Due to the complexity of the Information Extraction task in this challenging setting (sparse multi-class), the approaches that use the data annotated (e.g., (Arbabi et al., 2019; Kraljevic et al., 2019; Tiwari et al., 2020)) mostly rely on non-contextualised embeddings focusing on the detection precision. However, especially for clinical text, which is noisier and exhibits a variety of clinical expressions requiring disambiguation, relying on the context is essential. We argue that recall is very important, especially when automatic annotation results are used further in the downstream tasks.

In this work, we propose a self-supervised approach for sparse multi-class classification that fully relies on the context to detect contextual synonyms of medical concepts in clinical text. To be more precise, our model is based on the ClinicalBERT (Alsentzer et al., 2019) model which was pre-trained on the biomedical and clinical corpora that are widely used, producing state-of-the-art results in a range of supervised biomedical tasks, e.g. named entity recognition, relation extraction and question answering (Peng et al., 2019; Hahn and Oleynik, 2020). We separate the detection of frequent and rare classes by introducing different training objectives. The special training objective for rare classes increases the proximity of the respective textual embeddings and the ontology embeddings of concepts. Our work also exploits data augmentation techniques, such as paraphrasing and guided text generation to aid sparse class detection

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8754–8769
November 7–11, 2021.

and diversify the training data.

We apply our methodology for the phenotype detection task with more than 15,000 concepts from the Human Phenotype Ontology (HPO) (Köhler et al., 2017) [1]. The phenotyping task is an important Clinical NLP task that can improve the understanding of disease diagnosis (Aerts et al., 2006; Deisseroth et al., 2019; Liu et al., 2019a; Son et al., 2018; Xu et al., 2020). It remains unexplored due to the complexity of the classification into that large amount of classes. We test our approach on clinical data, namely on electronic health records (EHRs) and radiology reports.

Our **main contributions**: (1) Self-supervised methodology for contextual phenotype detection in clinical records. (2) Methodology for sparse class detection with the special training objective that increases proximity of contextual synonyms to ontology embeddings. (3) Data augmentation methodology to further improve the detection of sparse classes.

Our self-supervised models improve the current SOTA on F1 up to 4.5 absolute points, while on Recall up to 4.0 absolute points for the phenotype detection task for clinical data, which demonstrates how relying on the context is essential for this type of data. Second, after fine-tuning, our model outperforms the fine-tuned BERT-based models with as little as 20% of labelled data, which confirms efficiency of our self-supervised training objectives. Moreover, the extrinsic evaluation shows the benefits of using the phenotypes annotated by our model as features to predict ICU patient outcomes.

We present related work in Section 2, our phenotyping methods in Section 3, and our experimental setup in Section 4. Then, we present and discuss key results in Section 5. Finally, we conclude this work in Section 6.

2 Related Work

Most of the current methodologies for phenotype detection are supervised, BERT-based (e.g., BioBERT (Lee et al., 2019) or ClinicalBERT (Alsentzer et al., 2019)) and dedicated to the detection of certain rather limited phenotypes or their groups (Liu et al., 2019b; Zhang et al., 2019; Yang et al., 2020; Franz et al., 2020; Li et al., 2020).

Unsupervised methods in the clinical NLP domain traditionally rely on the usage of ontologies and knowledge bases. Human Phenotype Ontology (Köhler et al., 2017) is the most widely used ontology of phenotypes. The use of HPO in annotating phenotypic information automatically remains unexplored, mainly due to the complexity of formalising the task with over 15,000 concepts.

Such methods as MetaMap (Aronson and Lang, 2010) (the Mayo Clinic tool (Shen et al., 2017) based on it), cTAKES (Savova et al., 2010), NCBO (Jonquet et al., 2009) and ClinPhen (Deisseroth et al., 2019) follow similar pipelines and use linguistics patterns for shallow matching.

More recently, unsupervised deep learning methods have been applied to the problem, which allowed to perform the semantic analysis and go beyond shallow matching (Arbabi et al., 2019; Kraljevic et al., 2019; Tiwari et al., 2020). These approaches use non-contextualised embeddings, focus on the precision of detection with limited context exploitation. For example, the authors in (Kraljevic et al., 2019) propose a procedure to learn vectors of words enriched with their averaged context over the corpus to map them to correct medical concepts. We use contextualised word representations in contrast to all the related approaches and focus on recall.

3 Methodology

This section introduces the problem of phenotype detection along with our self-supervised method. It elaborates our data augmentation strategies, selective supervision in low-resource conditions, and finally explains our inference algorithm.

Problem Definition While annotating clinical text, clinicians usually relate HPOs to short spans, which usually have around 2-3 words depending on the corpus. [3] Following this rationale, we define the phenotype annotation as a two-step process: (1) detect HPO-relevant text spans, and (2) assign respective HPO concepts to those spans. More formally, given a textual document $X = \{t_1, ..., t_N\}$ represented by a sequence of tokens, and a full set of

[1] In the medical text, the word “phenotype” refers to deviations from normal morphology, physiology, or behaviour, such as skin rash, hypoxemia, neoplasm, etc. (Robinson, 2012). Note the difference of the phenotypic information to the diagnosis information expressed in ICD-10 codes (Organization, 2004). These codes record patient health states mainly for billing purposes. The former contributes to the latter.

[2] EMBL-EBI OLS: `https://www.ebi.ac.uk/ols/ontologies/hp`

[3] This general observation is confirmed in our internal annotation procedure (see Section 4.2)

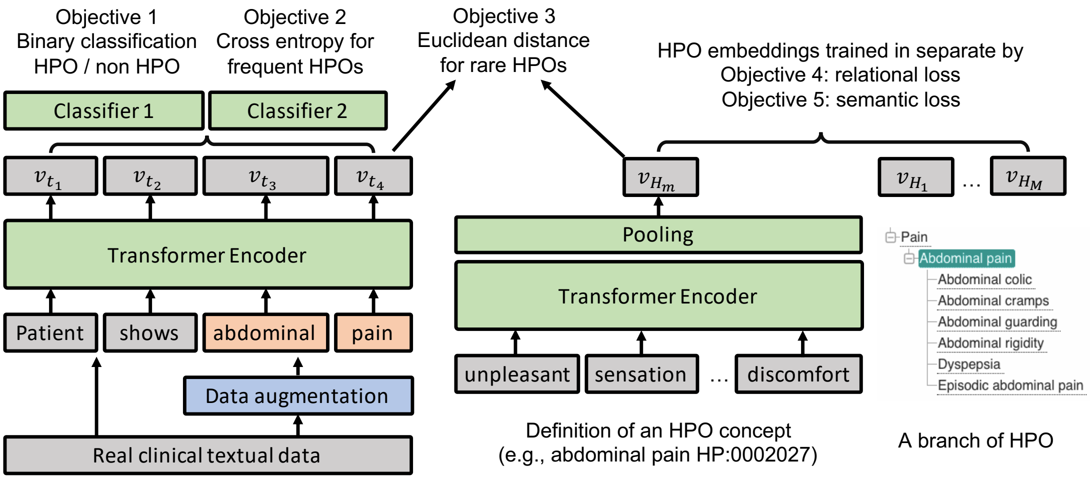

Figure 1: **Left**: Proposed model for phenotype annotation includes one Transformer encoder and two additional classifiers. The model is trained using the three self-supervised objectives (1, 2 and 3). **Middle and right**: Prior to the training of the phenotyping model, the other model for HPO embeddings is trained with relational and semantic losses (4, 5). Both Transformer encoders are initialised with ClinicalBERT but trained separately. The sub-figure of the HPO branch is taken from EMBL-EBI OLS [2].

HPO concepts $\mathcal{H} = \{H_1, ..., H_M\}$ under the root node *Phenotypic Abnormality (HP:0000118)* (exclusive) of the HPO ontology, our goal is to model: (1) $p(1_H|t_n)$, which is the conditional probability of the token t_n being HPO-relevant; (2) $p(H_m|t_n)$, which is the conditional probability if the HPO concept H_m should be assigned to the token t_n.

In the self-supervised setting, we consider only the training examples with textual spans matched with exact match to the HPO concepts as defined by the ontology. The main assumption here is that by capturing context of those term spans, the model will be able to generalise and detect formally different HPO spans seen in the similar contexts as the HPO concepts (*a.k.a.* contextual synonyms, for example, *Fever (HP:0001945)* will be matched to “feverish”). To support this challenging setting, we have designed a series of relevant training objectives described below.

Training Objectives As shown in Figure 1 (left), the proposed model for phenotype annotation consists of a Transformer encoder which is identical to and initialised with ClinicalBERT (Alsentzer et al., 2019). Besides, there are two additional classifiers on the top of the Transformer encoder which predict if a token is HPO-relevant and assign HPO concepts to those HPO-relevant tokens.

We enrich the model with the following three training objectives. (1) **A binary cross-entropy loss** $\mathcal{L}_1$ to predict $p(1_H|t_n)$, where 1_H is 1 if t_n is HPO-relevant, otherwise 0. (2) **A cross-entropy loss** $\mathcal{L}_2$ with softmax to predict $p(H_m|t_n)$, which is defined over the most frequent HPO concepts found in the training data. The intuition behind this objective is to increase precision of prediction in the resulting performance. (3) **The Euclidean distance** $\mathcal{L}_3$ between the token embedding v_{t_n} and the respective HPO concept embedding v_{H_m}, which is defined to increase recall of the model and targets the detection of the rare HPO concepts.

Note that the objectives above can be used for pre-training and further fine-tuning of the models in a way similar to BERT.

HPO Embeddings Prior to the training of the phenotyping model, we build the knowledge graph (KG) embeddings for HPO concepts. Figure 1 (middle and right) shows that this KG model has a Transformer encoder which learns the embeddings of HPO concepts given their definitions. It is designed to encode both the hierarchical connections between HPO concepts and the semantics in definitions of HPO concepts, so that the similar HPO concepts have similar embeddings. Therefore, we consider two learning objectives.

The first learning objective, namely **relational loss** $\mathcal{L}_4$, is to encourage the neighbouring HPO concepts to have similar embeddings and non-neighbouring HPO concepts to have different embeddings. The objective is implemented based on the distance of embeddings between neighbouring HPO concepts and non-neighbouring HPO concepts with softmax.

Second, the **semantic loss** $\mathcal{L}_5$ encourages the HPO embeddings to encode the semantics of input definitions and, more specifically, we adopt the skip-gram negative sampling (Mikolov et al., 2013).

3.1 Data Augmentation

There are two issues related to the creation of the training data by shallow matching: (1) this data can be too limited to help the model capture contextual phenotypes; (2) rare HPO concepts will not be found in the clinical text used for training and the model will not be able to detect them at the inference time. We are addressing those two problems by creating textual variants for existing HPO-relevant spans and generating context around rare HPO concepts.

HPO-relevant span variants with paraphrasing are used to replace the original spans in the training sentences. We create the variants by using the standard lexical pivoting paraphrasing technique where equivalent phrases in one language are found by "pivoting" over a shared translation into another language (Mallinson et al., 2017). We build the English-French-English pivot Seq2seq model.

Phenotypes are also often inferred from ranges of numerical values. E.g., *anemia (HP:0001903)* can be inferred from "Hgb 5 g/dl". We take the advantage of a series of reference laboratory values (from MIMIC (Johnson et al., 2016)) to create surrogates for the original names with numerical values. The named entities for which abnormal results are available are mapped to HPO concepts by an expert.

HPO context variants with synthetic text are created with a Seq2Seq model, which is trained to generate the textual context conditioned on HPO-relevant spans. For example, the sentence "patient was admitted with Angelman Syndrome to the ER" is generated given the input "Angelman Syndrome".

3.2 Decision Strategy for Inference

At the phenotype annotation inference stage, we assume that the HPO-relevant spans of frequent HPO concepts can be detected by $p(1_H|t_n)$ and $p(H_m|t_n)$ with high precision, while the Euclidean distance between contextualised token embedding v_{t_n} and HPO embedding v_{H_m} should be able to find those of rare HPO concepts with good recall. More precisely, we formalise the decision strategy as Algorithm 1.

Algorithm 1: The decision strategy of inferring phenotype annotation.

```
$X$ is the input sequence;
$\mathcal{H}$ stands for the full set of HPO concepts,
  $\mathcal{H}_{\text{freq}} \subset \mathcal{H}$ includes most frequent HPO
  concepts;
Initialise thresholds $\tau_p, \tau_d$ for $p(1_H|t_n)$ and
  distance function $D(v, u)$ respectively
  with pre-defined values;
for $t_n$ in $X = \{t_1, t_2, \dots t_N\}$ do
    if $p(1_H|t_n) \geq \tau_p$ then
        $r_n = \arg\max_{H_m} p(H_m|t_n)$ where
          $H_m \in \mathcal{H}_{\text{freq}}$ ;
    else if $\min_{H_m} D(v_{t_n}, v_{H_m}) < \tau_d$ then
        $r_n = \arg\min_{H_m} D(v_{t_n}, v_{H_m})$;
return $\{r_1, r_2, \dots, r_N\}$.
```

4 Experimental Setup

This section will introduce the datasets, implementation details, baselines and evaluation metrics.

4.1 Pre-training Corpora

EHR Corpus We use EHRs from the publicly available MIMIC-III database (Johnson et al., 2016). Diseases of the circulatory system are the most common reasons for those ICU stays. We collect the training samples from 38,772 notes of brief hospital course in MIMIC-III's discharge summaries and 1.5M generated notes by using data augmentation which is also trained on MIMIC-III.

Scientific Literature Corpus For the scientific text model, we use 119,924 PubMed abstracts (Cohan et al., 2018), $\sim$ 180k lines from the Cochrane data (Ive et al., 2016) and 1.5M generated notes by using data augmentation given PubMed abstracts.

Ontologies In the self-supervised setting, we consider HPO names, synonyms, abbreviations from the HPO as well as Unified Medical Language System (UMLS) (Bodenreider, 2004a) with exact match in clinical text as training samples.

4.2 Datasets

The following datasets are used as test data in the self-supervised setting, as well as train data in the supervised fine-tuning experiments.

	MIMIC	COVID-I	COVID-II	PubMed
#, articles	242	67	100	228
avg #, tokens	701.3	208.9	157.8	220.3
avg #, annotations	27.6	9.1	8.5	7.0
avg #, tok. / ann.	4.1	4.3	5.1	5.6
avg HPO depth	4.4	4.4	4.4	4.8
#, unique HPO	946	91	201	422
avg # ann. / HPO	7.0	6.6	4.2	3.8

Table 1: Statistics over the gold phenotype annotations of MIMIC, COVID-I, COVID-II, PubMed datasets. On average, each HPO appears less than 7.0 times.

Annotation Procedure To collect supervised datasets for evaluation and fine-tuning, we have annotated EHRs with HPO concepts with the help of three expert clinicians. The EHRs were pre-annotated with HPO concepts by keyword matching, and then the annotations were corrected by the three clinicians with consensus. The clinicians were specifically asked to identify **contextual synonyms** such as "drop in blood pressure" and "BP of 79/48" for *Hypotension (HP:0002615).*

MIMIC We have created our own sub-corpus of 242 discharge summaries from MIMIC-III with gold annotations. We used 146 EHRs for fine-tuning in the low-resource setting. 48 and 48 EHRs are reserved respectively for validation and testing in both self-supervised and supervised settings.

COVID We have collected and annotated two COVID datasets of short radiology reports: (1) **COVID-I** has 67 radiology reports from the Italian Society of Medical and Interventional Radiology [4] and (2) **COVID-II** is the International dataset with 100 radiology reports presented by (Cohen et al., 2020). From COVID-II, we have selected the patients with the diagnosis of the COVID-19 viral pneumonia. We take all the unique patients and extracted the longest (in terms of the tokens count) records for those patients. Reports from both datasets often contain not only the findings, but also the brief patient history. Both datasets are used as test sets for the self-supervised model. In the experiments with supervision, COVID-I was used to fine-tune and COVID-II to test.

PubMed To ensure comparison to the previous work, we also present our results for the PubMed dataset provided by (Groza et al., 2015) which contains 228 abstracts annotated by the creators of HPO. The common HPOs in this dataset are neurodevelopmental and skeletal disorders (e.g. Angelman syndrome), which is a quite different group of phenotypes as compared to the groups represented in the MIMIC and COVID data. An important difference between our annotation procedure as described above and the human annotation for the PubMed data is that the latter instructed annotating HPO-relevant spans only if they were presented in a canonical form close to HPO names: for example, "hypoplastic nails" and "nail hypoplasia" were included, but not "nails were hypoplastic". We re-use the random split: 40 abstracts for training and 188 for testing following NCR's setting (Arbabi et al., 2019). The statistics over the dataset is in Table 1.

4.3 Implementation Details [5]

The Transformer encoders in Figure 1 are initialised by ClinicalBERT, the two classifiers are two dense layers and the pooling layer concatenates max and average pooling. The maximum input length is 64 tokens. The proposed models are pre-trained for 100k steps and fine-tuned for 5k steps with batch size 64. The set of frequent HPO concepts $| \mathcal{H}_{\text{freq}} | = 400$ is decided by keyword matches. For data augmentation, we train a Seq2Seq Transformer model on a range of parallel English-French corpora in the biomedical field, namely the European Medicines Agency, Corpus of Parallel Patent Applications and the PatTR corpora.[6] The Seq2Seq model is based on OpenNMT (Klein et al., 2017). More details are given in Appendix C.

4.4 Setups

In the self-supervised setting, we train our models using either EHRs corpus for MIMIC and COVID (E) or scientific literature corpus for PubMed (S). We experiment with two setups with and without data augmentation.

We also evaluate the efficiency of our training objectives for pre-training and fine-tune our models with all the available supervised data.

However, in the real-life clinical setting, human annotations are very costly thus particular attention should be paid to the learning efficiency with a very small amount of data. We simulate this low-resource scenario and analyse the annotation cost / performance benefit trade-offs for our model. To

[4] `https://www.sirm.org/category/senza-categoria/covid-19`

[5] Due to the proprietary nature, the source code and gold annotations will not be shared publicly.

[6] `http://statmt.org/wmt14/medical-task/`

be more precise, we run a set of experiments where each time we pick a certain percentage of training examples according to one of the following strategies: (1) **Random sampling**: the samples are selected at random; (2) **Uncertainty-based sampling**: the entropy score based on $p(H_m|t_n), m \in \{1, 2, \ldots, M\}$ is computed to measure the uncertainty of the self-supervised model for each sample, and then the samples with the highest uncertainty score are selected; (3) **Oracle**: we also count the number of mismatched phenotypes between the keyword-based and gold annotations, and the samples with the most mismatches are selected.

4.5 Baselines

As baselines in the self-supervised setting, we report (1) Keyword: a naive method that simply matches HPO names, synonyms and abbreviations to text spans, (2) a range of text mining baselines (Clinphen (Deisseroth et al., 2019), NCBO (Jonquet et al., 2009), cTAKES (Savova et al., 2010), MetaMap (Aronson and Lang, 2010), MetaMapLite (Demner-Fushman et al., 2017)), and (3) two deep learning models (NCR (Arbabi et al., 2019), MedCAT (Kraljevic et al., 2019)) which are trained without supervision.

In the selective supervision setting, we use pre-trained models and fine-tune them on the datasets. More specifically, we use (1) BERT-Base (Devlin et al., 2019), (2) BioBERT-Base v1.0 (Lee et al., 2019) pre-trained on PubMed and PMC, (3) Clinical BERT (Alsentzer et al., 2019) pre-trained based on BioBERT and MIMIC-III discharge summaries, (4) SciBERT (Beltagy et al., 2019) pre-trained for scientific literature.

4.6 Metrics

We report the scores of micro-averaged Precision, Recall and F1-score at the document level. Following the best practices and to make our work comparable with others, we adopt the evaluation strategy of (Liu et al., 2019a). Thus, when we compute the following scores: (1) Exact match: only the exact same HPO annotations were counted as correct. (2) Generalised match: both the predicted and target HPO annotations are first extended to include all ancestors in HPO up until *Phenotypic Abnormality (HP:0000118)* (exclusive). Then the HPO annotations are de-duplicated for each document and the scores are computed.

5 Results and Discussion

This section discusses the results for the self-supervision and selective supervision settings.

Self-Supervised Setting We report results of the self-supervised model for the MIMIC, COVID, and PubMed datasets in Table 2. It compares the proposed model to the previous SOTA for the phenotyping task. Our principal observation is that our method outperforms all the baselines in terms of F1 and recall across datasets for both the exact and generalised matches. For example, for the exact match, our best models obtain F1 gain of at least 0.02, 0.05, and 0.02 and Recall scores gain of at least 0.04, 0.02, and 0.01 for MIMIC, COVID-I and COVID-II, respectively. This confirms the efficiency of our methodology for the detection of contextual synonyms in clinical text.

We note that our method does not give better performance for the PubMed dataset. We hypothesise that this happens due to the difference of gold annotation standards, as well as the fact that this dataset is oriented towards the detection of rare phenotypes with less frequent context patterns that are hence difficult to learn for our model.

Low-Resource Setting In this setting, we first study the efficiency of our self-supervised objectives for fine-tuning. Results are in Table 3 (more in Appendix B). Naturally fine-tuning leads to better automatic annotation accuracy on specific datasets. Our pre-training procedure is efficient and outperforms BERT-based models with at least 0.09, 0.16, 0.35 absolute increase in F1 (exact match) for the three datasets. Our analysis of the annotation cost / performance benefit trade-offs demonstrated that with only 20% of the training samples selected using the uncertainty criteria our fine-tuned model is able to achieve better F1 than ClinicalBERT which are fine-tuned on full training sets (see Figure 2).

The HPOs are sparse (less than 7 annotations on average) in the datasets as shown in Table 1. We further evaluate the model accuracy on annotating rare HPOs (any HPO excluding those from $\mathcal{H}_{\text{freq}}$ as defined in Section 4.3). Our fine-tuned model achieves F1 0.43 (exact match) and 0.60 (generalised match) on annotating rare HPOs while ClinicalBERT has F1 0.27 and 0.39 respectively and NCR achieves F1 0.34 and 0.47.

Qualitative Analysis To get better insights into the model performance, we have manually eye

Dataset	Method	Exact Match			Generalised Match		
		Precision	Recall	F1	Precision	Recall	F1
MIMIC	Keyword	0.7496	0.5223	0.6156	0.7883	0.6587	0.7177
	NCR	0.7747	0.4851	0.5967	0.8778	0.5917	0.7069
	NCBO	**0.9186**	0.3901	0.5477	**0.9632**	0.4711	0.6328
	Ours (E)	0.7334	**0.5619**	**0.6363**	0.7706	0.6972	0.7320
	Ours (E) w. Augmented Data	0.7235	0.5556	0.6285	0.7741	**0.6997**	**0.7351**
COVID-I	Keyword	0.6897	0.4710	0.5597	0.6750	0.5579	0.6109
	NCR	0.7873	0.4493	0.5721	0.8814	0.5481	0.6758
	NCBO	**0.8876**	0.4293	0.5788	**0.8857**	0.5070	0.6449
	Ours (E)	0.8617	0.4855	0.6211	0.8442	0.5657	0.6774
	Ours (E) w. Augmented Data	0.8576	**0.4909**	**0.6244**	0.8800	**0.5714**	**0.6929**
COVID-II	Keyword	0.8743	0.4514	0.5954	0.9268	0.5577	0.6963
	NCR	0.7220	0.4703	0.5696	0.9136	**0.6059**	**0.7286**
	NCBO	**0.9006**	0.4296	0.5817	**0.9484**	0.5128	0.6657
	Ours (E)	0.8517	**0.4811**	**0.6149**	0.9113	0.5814	0.7099
	Ours (E) w. Augmented Data	0.8421	0.4757	0.6079	0.8859	0.5695	0.6933
PubMed	Keyword	0.7221	0.5277	0.6098	0.8735	0.7175	0.7879
	NCR	0.7334	**0.6443**	**0.6860**	0.9131	**0.8183**	**0.8631**
	NCBO	**0.7948**	0.4441	0.5698	**0.9645**	0.6227	0.7568
	Ours (S)	0.6756	0.5121	0.5826	0.8741	0.7035	0.7796
	Ours (S) w. Augmented Data	0.6772	0.5627	0.6146	0.8818	0.7631	0.8182

Table 2: The proposed models in the self-supervised setting (without fine-tuning) achieved the best recall and F1 on MIMIC and COVID clinical text datasets. On PubMed which is scientific literature, our model clearly benefited from augmented data. Keyword, NCR and NCBO are reported as they achieve top F1 among the self-supervised baselines (Section 4.5) and full results are reported in Appendix A. The notations “Ours (E)” and “Ours (S)” refer to the models pre-trained on the EHR corpus and the scientific literature corpus, respectively.

Dataset	Method	Exact Match			Generalised Match		
		Precision	Recall	F1	Precision	Recall	F1
MIMIC	Fine-tuned ClinicalBERT	0.6962	0.5630	0.6225	0.8429	0.6980	0.7637
	Fine-tuned Ours (E) w. Augmented Data	**0.7141**	**0.7123**	**0.7132**	**0.8463**	**0.8380**	**0.8421**
COVID-II	Fine-tuned ClinicalBERT	0.6560	0.4138	0.5075	0.8063	0.5174	0.6303
	Fine-tuned Ours (E) w. Augmented Data	**0.7027**	**0.6324**	**0.6657**	**0.8652**	**0.7980**	**0.8302**
PubMed	Fine-tuned ClinicalBERT	0.5514	0.2449	0.3392	0.7715	0.4988	0.6059
	Fine-tuned Ours (S) w. Augmented Data	**0.7138**	**0.6618**	**0.6868**	**0.8959**	**0.8311**	**0.8623**

Table 3: The proposed model with fine-tuning in full achieved the best precision, recall and F1 scores on MIMIC, COVID-II and PubMed. The COVID-I is not reported as it is used to fine-tune the corresponding model. Only fine-tuned ClinicalBERT is reported as baseline because it achieves overall better F1 than fine-tuned BERT, BioBERT and SciBERT. Full results are available in Appendix B. The notations “Ours (E)” and “Ours (S)” refer to the models pre-trained on the EHR corpus and the scientific literature corpus, respectively.

Task (Metric)	Structured (Harutyunyan et al., 2019)	Structured	Structured + Phenotypes		
			+ NCR	+ ClinicalBERT	+ Ours
Length-of-stay (Kappa)	0.395	0.380	0.406	0.388	**0.430**
In-hospital Mortality (AUROC)	0.825	0.826	0.841	0.826	**0.845**
Decompensation (AUROC)	0.809	0.824	0.834	0.833	**0.839**

Table 4: Extrinsic evaluation on three ICU public benchmarks (Harutyunyan et al., 2019) which are created based on MIMIC-III. The results of (Harutyunyan et al., 2019) are reproduced by their code on the test set.

balled outputs of our MIMIC and COVID-I self-supervised model that achieves the best gain.

Our first observation is that our model is successful in capturing HPO-relevant contextual synonyms, which contributes to higher recall of our model. For example, “low pressure” and

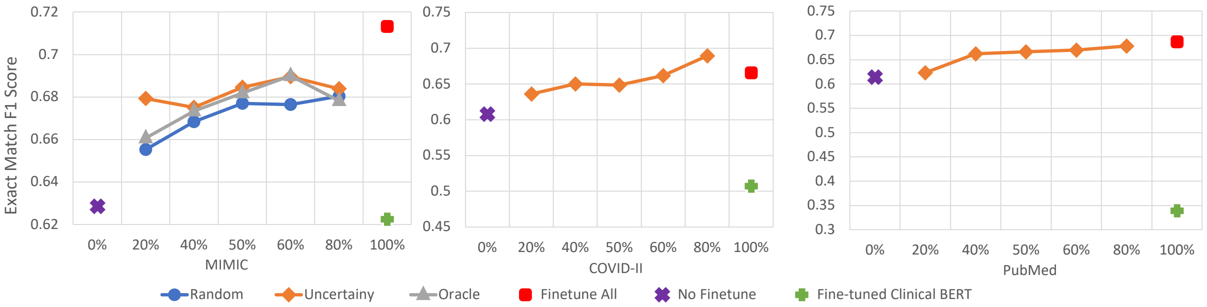

Figure 2: In the low resource setting with selective supervision, we pick subsets with 20%, 40%, 50%, 60%, 80% labelled data to fine-tune. The uncertainty sampling strategy is consistently better than the other two strategies on MIMIC and then applied on COVID-II and PubMed. The proposed models outperform fine-tuned BERT-based models with as little as 20% of labelled data. Details in Table 9 in Appendix B. Best to view in colours.

"hypotensive" are associated with *Hypotension (HP:0002615)* and "low platelets" with *Thrombocytopenia (HP:0001873)*.[7] Errors in the prediction mainly concern subtle distinctions between closely related phenotypes: *e.g.* "shortness of breath" triggers prediction *Respiratory Distress (HP:0002098)* whereas the gold label *Dyspnea (HP:0002094)* is the generalisation of *Respiratory Distress*.

For the more narrow-domain COVID-I dataset, false negatives often concern missed radiographic concepts: *e.g.* "perihilar infiltration" fails to trigger *Pulmonary Infiltrates (HP:0002113)*. In distinction to above, errors of our selective supervision models are less coupled with radiographic observations, e.g., false negatives for *Ankylosis (HP:0031013)*, *Abnormal Ear Morphology (HP:0031703)* and *Epileptic Spasm (HP:0011097)*.

Extrinsic Evaluation We evaluate the benefit of using phenotypes extracted by our models as features to enhance performance on downstream tasks. Following the setting by (Harutyunyan et al., 2019) with three public ICU benchmarks based on MIMIC-III, we train LSTMs with different input features: (1) 17 structured clinical features selected by (Harutyunyan et al., 2019) like heart rate and temperature or (2) structured clinical features plus phenotypes annotated by NCR, ClinicalBERT and our fine-tuned model respectively. The patients with both structured clinical features and textual notes are collected, and as a result, there are 21,346 patients (25,106 admissions) for training (with 4-fold cross validation) and 3,824 patients (4,497 admissions) for testing. Table 4 shows that the LSTMs which are fed with structured clinical features and phenotypes annotated by our model are consistently better than others on all three benchmarks. This demonstrates that increasing recall in phenotyping is essential for downstream tasks.

6 Conclusion

In this paper, we have proposed a deep self-supervised phenotype annotation approach relying on contextualised word embeddings and data augmentation techniques. Our experimental results in a challenging sparse multi-class setting, with over 15,000 candidate HPO concepts, indicate that our methodology is particularly efficient to detect contextual mentions of phenotype concepts in clinical text. We demonstrate that increasing phenotyping recall is essential for downstream tasks.

7 Ethics Considerations

The study has been carried out in accordance with relevant guidelines and regulations for the MIMIC-III data. Other data used in this study can be accessed without any preliminary requests. Clinical experts received consulting fees for their work. The purpose of the developed models is to extract phenotypic information from unstructured healthcare data. This information is only to assist human medical experts in their decisions. Before the deployment in the actual clinical setting our methodology is subject to systematic debugging, extensive simulation, testing and validation under the supervision of expert clinicians.

8 Acknowledgement

We would like to thank Dr. Garima Gupta, Dr. Deepa (M.R.S.H) and Dr. Ashok (M.S.) for helping us create gold-standard phenotype annotation data.

[7] All examples hereinafter are paraphrased.

We would also like to thank the four anonymous reviewers for the feedback.

References

Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. 2006. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aryan Arbabi, David R Adams, Sanja Fidler, and Michael Brudno. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7(2):e12596.

Alan R Aronson. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004a. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–70.

Olivier Bodenreider. 2004b. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:267–270.

Thomas A Caswell, Michael Droettboom, Antony Lee, John Hunter, Elliott Sales de Andrade, Eric Firing, Tim Hoffmann, Jody Klymak, David Stansby, Nelle Varoquaux, Jens Hedegaard Nielsen, Benjamin Root, Ryan May, Phil Elson, Jouni K. Seppänen, Darren Dale, Jae-Joon Lee, Damon McDougall, Andrew Straw, Paul Hobson, Christoph Gohlke, Tony S Yu, Eric Ma, Adrien F. Vincent, Steven Silvester, Charlie Moad, hannah, Nikita Kniazev, Elan Ernest, and Paul Ivanov. 2020. matplotlib/matplotlib: REL: v3.3.3.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. 2020. Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*.

Cole A Deisseroth, Johannes Birgmeier, Ethan E Bodle, Jennefer N Kohler, Dena R Matalon, Yelena Nazarenko, Casie A Genetti, Catherine A Brownstein, Klaus Schmitz-Abe, Kelly Schoch, et al. 2019. Clinphen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genetics in Medicine*, 21(7):1585–1593.

Dina Demner-Fushman, Willie J Rogers, and Alan R Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Leopold Franz, Yash Raj Shrestha, and Bibek Paudel. 2020. A deep learning pipeline for patient diagnosis prediction using electronic health records. *arXiv preprint arXiv:2006.16926*.

Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N Robinson. 2015. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015.

Udo Hahn and Michel Oleynik. 2020. Medical Information Extraction in the Age of Deep Learning. *Yearbook of medical informatics*, 29(1):208–220.

Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren

Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18.

Irvin, Rajpurkar, Ko, Yu, Ciurea-Ilcus, Chute, Marklund, Haghgoo, Ball, and Shpanskaya. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Julia Ive, AurÉlien Max, François Yvon, and Philippe Ravaud. 2016. Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations. In *International Conference on Language Resources and Evaluation - Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016 2016)*, page 8, Portorož, Slovenia.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Clement Jonquet, Nigam Shah, Cherie Youn, Chris Callendar, Margaret-Anne Storey, and M Musen. 2009. Ncbo annotator: semantic annotation of biomedical data. In *International Semantic Web Conference, Poster and Demo session*, volume 110.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, et al. 2017. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876.

Zeljko Kraljevic, Daniel Bean, Aurelie Mascio, Lukasz Roguski, Amos Folarin, Angus Roberts, Rebecca Bendayan, and Richard Dobson. 2019. Medcat – medical concept annotation tool.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12.

Cong Liu, Casey N Ta, James R Rogers, Ziran Li, Junghwan Lee, Alex M Butler, Ning Shang, Fabricio Sampaio Peres Kury, Liwei Wang, Feichen Shen, Hongfang Liu, Lyudmila Ena, Carol Friedman, and Chunhua Weng. 2019a. Ensembles of natural language processing systems for portable phenotyping solutions. *Journal of Biomedical Informatics*, 100:103318.

Dianbo Liu, Dmitriy Dligach, and Timothy Miller. 2019b. Two-stage federated phenotyping and patient representation learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 283–291, Florence, Italy. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

World Health Organization. 2004. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, Wang, Lu, Bagheri, Summers, and Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Peter N Robinson. 2012. Deep phenotyping for precision medicine. *Human mutation*, 33(5):777–780.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):507–513.

Feichen Shen, Liwei Wang, and Hongfang Liu. 2017. Phenotypic Analysis of Clinical Narratives Using Human Phenotype Ontology. *Studies in health technology and informatics*, 245:581–585.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Jung Hoon Son, Gangcai Xie, Chi Yuan, Lyudmila Ena, Ziran Li, Andrew Goldstein, Lulin Huang, Liwei Wang, Feichen Shen, Hongfang Liu, et al. 2018. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *The American Journal of Human Genetics*, 103(1):58–73.

Prayag Tiwari, Sagar Uprety, Shahram Dehdashti, and M Shamim Hossain. 2020. TermInformer: unsupervised term mining and analysis in biomedical literature. *Neural Computing and Applications*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhenxing Xu, Jingyuan Chou, Xi Sheryl Zhang, Yuan Luo, Tamara Isakova, Prakash Adekkanattu, Jessica S Ancker, Guoqian Jiang, Richard C Kiefer, Jennifer A Pacheco, Luke V Rasmussen, Jyotishman Pathak, and Fei Wang. 2020. Identifying subphenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. *Journal of biomedical informatics*, 102:103361.

Zhen Yang, Matthias Dehmer, Olli Yli-Harja, and Frank Emmert-Streib. 2020. Combining deep learning with token selection for patient phenotyping from electronic health records. *Scientific Reports*, 10(1):1432.

Jingqing Zhang, Xiaoyu Zhang, Kai Sun, Xian Yang, Chengliang Dai, and Yike Guo. 2019. Unsupervised annotation of phenotypic abnormalities via semantic latent representations on electronic health records. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 598–603. IEEE.

A Results of the Self-Supervised Models

MIMIC						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
Keyword	0.7496	0.5223	0.6156	0.7883	0.6587	0.7177
NCR	0.7747	0.4851	0.5967	0.8778	0.5917	0.7069
Clinphen	0.8147	0.3148	0.4541	0.9337	0.4066	0.5665
NCBO	**0.9186**	0.3901	0.5477	**0.9632**	0.4711	0.6328
cTAKES	0.8250	0.3259	0.4673	0.9321	0.4287	0.5872
MetaMap	0.7909	0.4062	0.5367	0.8835	0.5384	0.6691
MetaMapLite	0.7968	0.4358	0.5634	0.8766	0.5720	0.6923
MedCAT (Medmentions)	0.7290	0.3321	0.4563	0.8305	0.4711	0.6012
MedCAT (UMLS)	0.8630	0.3889	0.5362	0.9311	0.5231	0.6699
Ours (E)	0.7334	**0.5619**	**0.6363**	0.7706	0.6972	0.7320
Ours (E) w. Augmented Data	0.7235	0.5556	0.6285	0.7741	**0.6997**	**0.7351**

Table 5: Results on MIMIC in the self-supervised setting.

COVID-I						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
Keyword	0.6897	0.4710	0.5597	0.6750	0.5579	0.6109
NCR	0.7873	0.4493	0.5721	0.8814	0.5481	0.6758
Clinphen	0.8259	0.3351	0.4768	0.8693	0.4042	0.5518
NCBO	**0.8876**	0.4293	0.5788	**0.8857**	0.5070	0.6449
cTAKES	0.7305	0.1866	0.2973	0.8285	0.3112	0.4524
MetaMap	0.8023	0.3750	0.5111	0.8990	0.5039	0.6458
MetaMapLite	0.7765	0.3587	0.4907	0.8992	0.4914	0.6355
MedCAT (Medmentions)	0.7519	0.3514	0.4790	0.8284	0.4940	0.6189
MedCAT (UMLS)	0.6293	0.2645	0.3724	0.8295	0.4171	0.5551
Ours (E)	0.8617	0.4855	0.6211	0.8442	0.5657	0.6774
Ours (E) w. Augmented Data	0.8576	**0.4909**	**0.6244**	0.8800	**0.5714**	**0.6929**
COVID-II						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
Keyword	0.8743	0.4514	0.5954	0.9268	0.5577	0.6963
NCR	0.7220	0.4703	0.5696	0.9136	**0.6059**	**0.7286**
Clinphen	0.7789	0.3290	0.4626	0.9038	0.4256	0.5787
NCBO	**0.9006**	0.4296	0.5817	**0.9484**	0.5128	0.6657
cTAKES	0.7684	0.2098	0.3296	0.9158	0.3327	0.4881
MetaMap	0.8517	0.3218	0.4672	0.9437	0.4152	0.5767
MetaMapLite	0.7828	0.3261	0.4604	0.9494	0.4431	0.6042
MedCAT (Medmentions)	0.7599	0.3046	0.4349	0.8757	0.4284	0.5753
MedCAT (UMLS)	0.8333	0.2586	0.3947	0.9368	0.3675	0.5280
Ours (E)	0.8517	**0.4811**	**0.6149**	0.9113	0.5814	0.7099
Ours (E) w. Augmented Data	0.8421	0.4757	0.6079	0.8859	0.5695	0.6933

Table 6: Results on COVID-I and COVID-II in the self-supervised setting.

PubMed						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
Keyword	0.7221	0.5277	0.6098	0.8735	0.7175	0.7879
NCR	0.7334	**0.6443**	**0.6860**	0.9131	**0.8183**	**0.8631**
Clinphen	0.6352	0.3926	0.4853	0.9240	0.5095	0.6568
NCBO	**0.7948**	0.4441	0.5698	**0.9645**	0.6227	0.7568
cTAKES	0.5602	0.2216	0.3175	0.8953	0.3479	0.5011
MetaMap	0.7167	0.4966	0.5867	0.9076	0.6671	0.7690
MetaMapLite	0.7057	0.4334	0.5370	0.8978	0.5934	0.7146
MedCAT (Medmentions)	0.5362	0.2089	0.3007	0.7387	0.3066	0.4333
MedCAT (UMLS)	0.7636	0.4237	0.5450	0.9376	0.5903	0.7245
Ours (S)	0.6756	0.5121	0.5826	0.8741	0.7035	0.7796
Ours (S) w. Augmented Data	0.6772	0.5627	0.6146	0.8818	0.7631	0.8182

Table 7: Results on PubMed in the self-supervised setting.

B Selective Supervision Results in Low-Resource Setting

MIMIC						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
BERT	0.7132	0.5617	0.6285	0.8434	0.6844	0.7557
BioBERT	0.6864	0.5728	0.6245	0.8335	0.7021	0.7622
ClinicalBERT	0.6962	0.5630	0.6225	0.8429	0.6980	0.7637
SciBERT	0.6898	0.5407	0.6062	0.8269	0.6671	0.7385
Ours (E) w. Augmented Data	**0.7141**	**0.7123**	**0.7132**	**0.8463**	**0.8380**	**0.8421**
COVID-II						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
BERT	0.6144	0.3549	0.4499	0.8193	0.4760	0.6022
BioBERT	0.5858	0.3922	0.4699	0.7781	0.5201	0.6235
ClinicalBERT	0.5711	0.4095	0.4770	0.7680	0.5039	0.6085
SciBERT	0.6560	0.4138	0.5075	0.8063	0.5174	0.6303
Ours (E) w. Augmented Data	**0.7027**	**0.6324**	**0.6657**	**0.8652**	**0.7980**	**0.8302**
PubMed						
Method	Exact Match			Generalised Match		
	Precision	Recall	F1	Precision	Recall	F1
BERT	0.5103	0.2400	0.3265	0.7530	0.4795	0.5859
BioBERT	0.4828	0.2459	0.3258	0.7716	0.4911	0.6002
ClinicalBERT	0.5514	0.2449	0.3392	0.7715	0.4988	0.6059
SciBERT	0.4967	0.2177	0.3027	0.7187	0.4638	0.5638
Ours (E) w. Augmented Data	**0.7138**	**0.6618**	**0.6868**	**0.8959**	**0.8311**	**0.8623**

Table 8: Results on MIMIC, COVID-II and PubMed with supervision. All models are fine-tuned with full training samples.

MIMIC							
Sampling Strategy	Ratio	Exact Match			Generalised Match		
		Precision	Recall	F1	Precision	Recall	F1
Random	20%	0.6496	0.6609	0.6552	0.7760	0.8074	0.7914
	40%	0.6441	0.6943	0.6683	0.7709	0.8408	0.8044
	50%	0.6529	0.7030	0.6770	0.7755	0.8357	0.8045
	60%	0.6457	0.7104	0.6765	0.7744	0.8367	0.8044
	80%	0.6508	0.7129	0.6804	0.7765	0.8419	0.8078
Oracle	20%	0.6344	0.6894	0.6607	0.7521	0.8200	0.7846
	40%	0.6349	0.7166	0.6733	0.7631	0.8453	0.8021
	50%	0.6422	0.7265	0.6818	0.7562	0.8575	0.8037
	60%	0.6530	0.7314	0.6900	0.7626	0.8582	0.8076
	80%	0.6370	0.7252	0.6782	0.7707	0.8524	0.8095
Uncertainty	20%	0.6707	0.6881	0.6793	0.7667	0.8142	0.7898
	40%	0.6592	0.6918	0.6751	0.7692	0.8269	0.7970
	50%	0.6503	0.7228	0.6846	0.7659	0.8476	0.8047
	60%	0.6580	0.7240	0.6895	0.7737	0.8541	0.8119
	80%	0.6499	0.7215	0.6839	0.7722	0.8514	0.8099
COVID-II							
Sampling Strategy	Ratio	Exact Match			Generalised Match		
		Precision	Recall	F1	Precision	Recall	F1
Uncertainty	20%	0.6990	0.5838	0.6362	0.8680	0.7261	0.7907
	40%	0.7129	0.5973	0.6500	0.8796	0.7303	0.7980
	50%	0.6978	0.6054	0.6483	0.8719	0.7610	0.8127
	60%	0.7220	0.6108	0.6618	0.8874	0.7603	0.8190
	80%	0.7573	0.6324	0.6892	0.8951	0.7750	0.8307
PubMed							
Sampling Strategy	Ratio	Exact Match			Generalised Match		
		Precision	Recall	F1	Precision	Recall	F1
Uncertainty	20%	0.6899	0.5685	0.6233	0.8852	0.7766	0.8273
	40%	0.7156	0.6161	0.6621	0.8922	0.7949	0.8408
	50%	0.6877	0.6463	0.6663	0.8798	0.8226	0.8502
	60%	0.7014	0.6414	0.6701	0.8883	0.8141	0.8496
	80%	0.7028	0.6550	0.6781	0.8867	0.8192	0.8516

Table 9: Results on MIMIC, COVID-II and PubMed in the selective supervision setting. The F1 scores of exact match correspond to Figure 2.

C Implementation Details

The data processing and model are developed by Python 3.6. Besides our own code, we use open-sourced third-party libraries including Matplotlib (Caswell et al., 2020), Numpy (Harris et al., 2020), Pandas, Pronto, Scikit-learn (Pedregosa et al., 2011), Transformers (Wolf et al., 2020), Tensorboard, Pytorch (Paszke et al., 2019) (v1.7, CUDA 10.1), Tqdm, Xmltodict. The number of learnable parameters is close to a BERT-base model. On two NVIDIA TITANX GPUs, it takes around 24 hours to pre-train and 1.2 hours to fine-tune.

Self-supervised training	
Optimiser	AdamW
Training steps	100k
Learning rate	1e-4
Batch size	64
Vocab size	28996
Maximum input length	64
Fine-tuning	
Optimiser	AdamW
Training steps	5k
Learning rate	1e-4
Batch size	64
Vocab size	28996
Maximum input length	64
HPO Embeddings	
Optimiser	AdamW
Training steps	30k
Learning rate	2e-5
Batch size	64
Vocab size	28996
Maximum input length	64

Table 10: Hyper-parameters for training are decided empirically on the validation set.