Phenotyping in clinical text with unsupervised numerical reasoning for patient stratification

Ashwani Tanwar*, Jingqing Zhang*, Julia Ive, Vibhor Gupta and Yike Guo

Pangaea Data Limited, London SE1 7LY, UK
*These authors contributed equally to this paper.
Corresponding author: Yike Guo. Email: yguo@pangaeadata.ai

Impact Statement

Profiling a patient using the phenotypes from clinical text has various applications in the healthcare domain, including patient stratification for hard-to-diagnose diseases. One of the key challenges in phenotyping is numerical reasoning which involves determining a phenotype based on numerical measurements. This is an open problem, and current state-of-the-art generic phenotyping methods perform poorly. Our study addresses this issue by presenting a novel unsupervised methodology that substantially outperforms the generic phenotyping methods. Our methodology is helpful for our research community as it requires no supervision, saving a lot of cost and time, as well as it can be generalized to other biomedical tasks requiring numerical reasoning. We demonstrate the impact of the work by building patient stratification systems for several diseases by using these phenotypes which can improve accuracy and substantially lessen the manual effort and time involved in screening of patients. In addition, the work provides a new solution to improve clinical data quality by imputing missing values into structured data.

Abstract

Phenotypic information of patients, as expressed in clinical text, is important in many clinical applications such as identifying patients at risk of hard-to-diagnose conditions. Extracting and inferring some phenotypes from clinical text requires numerical reasoning, for example, a temperature of 102°F suggests the phenotype Fever. However, while current state-of-the-art phenotyping models using natural language processing (NLP) are in general very efficient in extracting phenotypes, they struggle to extract phenotypes that require numerical reasoning. In this article, we propose a novel unsupervised method that leverages external clinical knowledge and contextualized word embeddings by ClinicalBERT for numerical reasoning in different phenotypic contexts. Experiments show that the proposed method achieves significant improvement against unsupervised baseline methods with absolute increase in generalized Recall and F1 scores of up to 79% and 71%, respectively. Also, the proposed method outperforms supervised baseline methods with absolute increase in generalized Recall and F1 scores of up to 70% and 44%, respectively. In addition, we validate the methodology on clinical use cases where the detected phenotypes significantly contribute to patient stratification systems for a set of diseases, namely, HIV and myocardial infarction (heart attack). Moreover, we find that these phenotypes from clinical text can be used to impute the missing values in structured data, which enrich and improve data quality.

Keywords: Numerical reasoning, phenotyping, contextualized word embeddings, patient stratification, unsupervised learning, natural language processing, deep learning

***Experimental Biology and Medicine* 2022; 247: 2038–2052. DOI: 10.1177/15353702221118092**

Introduction

Phenotype extraction from unstructured clinical text is shown to be useful for various clinical applications[1] such as predicting intensive care unit (ICU) in-hospital mortality, predicting length of stay of patients, and identifying patients with hard-to-diagnose diseases. In this study, the word "phenotype" refers to deviations from normal morphology, physiology, or behavior, such as skin rash, hypoxemia, and neoplasm.[2] Please note the difference in the phenotypic information from the diagnosis information expressed in International Classification of Disease (ICD) codes,[3] as the former contributes to the latter. One challenge in extracting phenotypes is numerical reasoning (NR), as many of the phenotypes rely on bedside measurements such as temperature, heart rate, breathing rate, blood pressure, serum creatinine, hematocrit, and glucose levels, which we refer to as numeric entities. As these phenotypes require reasoning with the numbers based on clinical knowledge, they are often missed or incorrectly extracted by the existing phenotyping methods that are not designed to consider NR.[4–10]

Current phenotyping methods based on state-of-the-art (SOTA) machine learning (ML) and natural language processing (NLP) techniques mostly exploit non-contextualized word embeddings. For example, neural concept recognizer (NCR)[8] utilizes convolutional neural networks (CNNs) to

ISSN 1535-3702

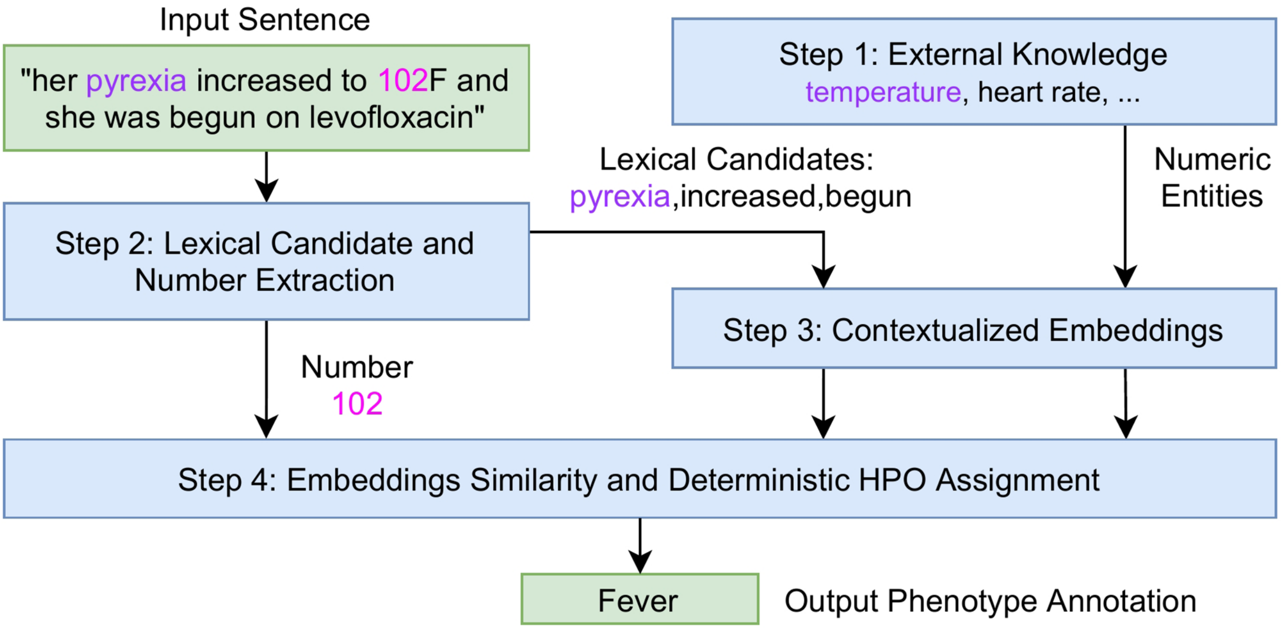

Input Sentence
"her pyrexia increased to 102F and she was begun on levofloxacin"
Step 1: External Knowledge
temperature, heart rate, ...
Lexical Candidates:
pyrexia,increased,begun
Numeric Entities
Step 2: Lexical Candidate and Number Extraction
Step 3: Contextualized Embeddings
Number
102
Step 4: Embeddings Similarity and Deterministic HPO Assignment
Fever
Output Phenotype Annotation

Figure 1. The proposed method of numerical reasoning (NR) model to extract specific phenotypes from clinical text without supervision.

build non-contextualized embeddings for biomedical concepts defined by ontologies such as human phenotype ontology (HPO).[11] These methods are limited to detect contextual synonyms of phenotypes, as clinicians may describe the same phenotype in a variety of ways in clinical text. For example, previous SOTA phenotyping models such as NCR and NCBO[6] fail to capture the phenotype *Fever* from the sentence "patient is reported to have high temperature." The recent study[12] extended the above works to use contextualized embeddings (Bidirectional Encoder Representations from Transformers [BERT]-based[13]) to capture phenotypes from different contexts.

However, none of these methods mentioned above can reason with numbers in clinical text, for example, "temperature 102°F" suggesting *Fever*. Recent works in NR publish new datasets[14] and develop NR capabilities with deep learning[15–18] in the respective domains rather than the clinical domain. For instance, Geva *et al.*[19] show better performance on tasks involving numeracy, such as math word problems, and reading comprehension by using artificially created data. Other works[20–22] introduce extra NR modules into deep learning, which are very specific to the numeracy tasks such as a calculator for arithmetic operations and thus cannot be generalized to reason in the clinical context. Overall, although these models show benefits in their respective domains, they did not incorporate clinical knowledge to address challenges in clinical applications.[23]

In practice, NR for clinical context has the following challenges. First, there can be accumulation of multiple numeric examples in a condensed context, such as "Physical examination: temperature 97.5, blood pressure 124/55, pulse 79, respirations 18, O_2 saturation 99% on room air." Second, the contexts of numeric examples can be different, such as "temperature of 102°F," "temperature is 102°F," "temperature is recorded as 102°F," "temperature is found to be 102°F," which require more robust models to identify the (*numeric entity, number*) pair, namely, *temperature, 102°F* in this case. Third, not all numbers in clinical text are connected with phenotypes. For instance, the number in "patient required 4 days of hospitalization" is not related with any phenotype. We aim to address all the three challenges using our novel NR methodology.

To the best of our knowledge, previous studies have not addressed these challenges and this article proposes a novel deep learning-based (BERT-based) unsupervised method to accurately extract phenotypes with NR from various clinical contexts by leveraging clinical external knowledge. In summary, our main contributions are as follows:

1. We propose a novel unsupervised method to accurately extract phenotypes that require NR by using NLP and deep learning techniques. The proposed method can extract phenotypes from various contexts with contextualized word embeddings.
2. Intrinsic evaluation shows significant superior accuracy of the proposed method against alternative phenotyping methods in both unsupervised and supervised settings.
3. Extrinsic evaluation shows the contributions of the phenotypes extracted by the proposed method to support better patient stratification. Also, these phenotypes are shown to impute and enrich the missing structured data, which improves data quality for potential downstream applications.

Materials and methods

Figure 1 presents the proposed unsupervised method for NR to extract specific phenotypes (defined as HPO[24]) from clinical notes. The proposed method consists of four steps, which will be elaborated upon in this section: (1) an external one-time knowledge base is created which connects numeric

Table 1. Numeric entities with normal reference range.

ID	Numeric entity	Abbreviation	Unit	Normal reference range	
				Lower bound	Upper bound
0	Temperature	Temp	Celsius	36.4	37.3
0	Temperature	Temp	Fahrenheit	97.5	99.1
1	Heart rate	Heart rate	Beats per minute (bpm)	60	80
2	Breathing rate	Breathing rate	Breaths per minute	12	20
3	Serum creatinine	Serum creatinine	mg/dL	0.6	1.2
3	Serum creatinine	Serum creatinine	micromoles/L	53	106.1
4	Hematocrit	Hct	%	41	48
5	Blood oxygen	O_2	%	95	100

Examples of numeric entities with their corresponding normal reference range and units. A complete table of all numeric entities considered by this study is provided in Table 14.

Table 2. Numeric entities and corresponding phenotypes.

ID	Numeric entity	Number lower than the lower bound (affirmed)		Number higher than the upper bound (affirmed)		Number inside normal range (negated)	
		HPO ID	HPO Name	HPO ID	HPO Name	HPO ID	HPO name
0	Temperature	HP:0002045	Hypothermia	HP:0001945	Fever	HP:0004370	Abnormality of temperature regulation
1	Heart rate	HP:0001662	Bradycardia	HP:0001649	Tachycardia	HP:0011675	Arrhythmia
2	Breathing rate	HP:0046507	Bradypnea	HP:0002789	Tachypnea	HP:0002793	Abnormal pattern of respiration
3	Serum creatinine	HP:0012101	Decreased serum creatinine	HP:0003259	Elevated serum creatinine	HP:0012100	Abnormal circulating creatinine concentration
4	Hematocrit	HP:0031851	Reduced hematocrit	HP:0001899	Increased hematocrit	HP:0031850	Abnormal hematocrit
5	Blood oxygen	HP:0012418	Hypoxemia	HP:0012419	Hyperoxemia	HP:0500165	Abnormal blood oxygen level

HPO: human phenotype ontology.
Examples of numeric entities and their corresponding phenotype labels (formalized by HPO ID and HPO name). The ID column links with Table 1. A complete table of all numeric entities considered by this study is provided in Table 15.

entities and phenotypes, (2) numbers and lexical candidates for numeric entities are then extracted from input text, (3) contextualized embeddings for numeric entities and lexical candidates are computed, respectively, and (4) the output phenotype is determined based on embedding similarity between lexical candidates and numeric entities.

External knowledge

Specific phenotypes can often be inferred from clinical text by numbers and numeric entities. For instance, in clinical notes, the numeric entity "temperature" and the numerical value "102 Fahrenheit" together suggest the phenotype *Fever (HP:0001945)* of a patient. Thus, prior to other steps, an external knowledge base is created to connect phenotypes, numeric entities, and numerical values.

To connect numeric entities and numerical values, as shown in Table 1, we manually collect the most frequent numeric entities such as temperature, heart rate, breathing rate, serum anion gap, and platelet and their normal reference ranges (values and units) from the National Health Service (NHS) of UK (Accessed in April 2022: https://www.nhs.uk) and MIMIC-III.[25] For example, we use the NHS website to search for "temperature" and then specifically look for the numeric entities, normal (healthy) reference ranges, and the corresponding unit. We find the lower bound of normal temperature is 97.5°F (or 36.4°C) while the upper bound is 99.1°F (or 37.3°C), which are then validated by three expert clinicians with consensus.

Second, Table 2 connects numeric entities and numerical values with phenotypes (i.e. HPO concepts). A numeric entity is typically connected with three phenotypes representing when a measurement reading is lower than, higher than, or within the normal reference range of numerical values. For example, when a temperature reading is lower or higher than the normal range, a patient is affirmed to have the phenotype *hypothermia* (*HP:0002045*) or *fever* (*HP:0001945*), respectively. Otherwise, a patient is negated to have *abnormality of temperature regulation* (*HP:0004370*), which is the parent phenotype of *hypothermia* (*HP:0002045*) or *fever* (*HP:0001945*). Both Tables 1 and 2 are validated by three expert clinicians with consensus for authenticity.

Number and lexical candidate extraction

Given clinical text, numbers and lexical candidates which are likely to be numeric entities are extracted, as shown in Figure 2. More specifically, regular expression is first used to extract numbers in alphanumeric format like "pyrexia increased to 102°F" and "heart rate in 90s," while dates or numbers in clinical concepts like "vitamin B12" and "O_2 saturation" are excluded with a predefined dictionary.[26] Then, syntactic analysis is used to extract the lexical candidates (with focus on nouns, adjectives, and verbs) that are linked

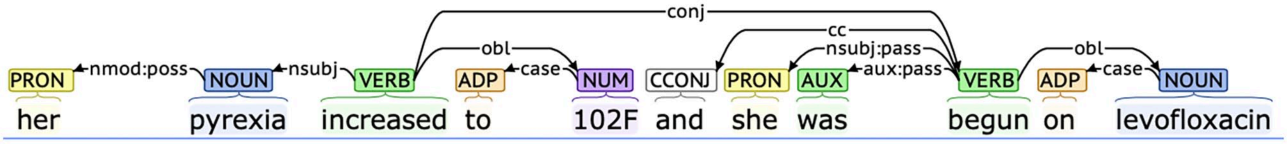


Figure 2. An example to extract lexical candidates of numbers by syntactic analysis. In this example, the words "pyrexia," "increased," "begun" are extracted as the lexical candidates of the number "102°F" by using part-of-speech (POS) tagging and dependency parsing.

with these numbers. For instance, Figure 2 shows that the words "pyrexia," "increased," and "begun" are considered as lexical candidates because they are linked to the number "102°F" within two hops, which suggest they are likely to be numeric entities. The extraction of lexical candidates encourages more extractions so that no important candidate is missed, but not all of the lexical candidates will eventually be considered as a numeric entity in later steps.

Contextualized embeddings for numeric entities and lexical candidates

To decide whether a lexical candidate is a numeric entity, we first create contextualized embeddings for numeric entities and lexical candidates by fine-tuning ClinicalBERT[27] and then measure their similarity. To create contextualized embeddings of numeric entities, the objective is to create a semantic embedding space in which the embeddings of all possible expressions (names and synonyms) of the same numeric entity are clustered closer, while those with different numeric entities are differentiated. Therefore, ClinicalBERT is fine-tuned with semantic textual similarity (STS) as in the equation below, which aims to maximize the cosine similarity between expressions of the same numeric entity while it minimizes the similarity between different numeric entities

$$\mathcal{L}\left(e_i,s_j\right)=\frac{1}{|\mathcal{E}|}\frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{E}|}\sum_{j=1}^{|\mathcal{S}|}\left(\cos\left(\mathbf{h}_{e_i},\mathbf{h}_{s_j}\right)-y_{e_i,s_j}\right)^2$$
$$\text{where } y_{e_i,s_j}=\begin{cases}1, & \text{if } s_j \text{ is a synonym of } e_i\\ 0, & \text{otherwise}\end{cases} \qquad (1)$$

where $\mathbf{h}_{e_i}$ is the contextualized embedding of the ith numeric entity e_i in $\mathcal{E}$. Similarly, $\mathbf{h}_{s_j}$ is the contextualized embedding for the jth synonym s_j in $\mathcal{S}$. To collect the training data for STS, the numeric entities $\mathcal{E}=\{e_1,e_2,\ldots,e_{|\mathcal{E}|}\}$ are listed in Table 2 and their corresponding synonyms $\mathcal{S}=\{s_1,s_2,\ldots,s_{|\mathcal{S}|}\}$ are collected from HPO and the unified medical language system (UMLS).[28] During inference, the contextualized embeddings of lexical candidates are computed by feeding the lexical candidates into the same fine-tuned ClinicalBERT model on-the-fly.

Embedding similarity and deterministic HPO assignment

After the creation and computation of contextualized embeddings of numeric entities and lexical candidates, embedding pairs by Cartesian product between lexical candidates and numeric entities are created and then cosine similarity is calculated between all the pairs. The pair with the maximum cosine similarity above a predefined threshold connects the lexical candidate with its corresponding numeric entity.

Once the numeric entity is decided, the phenotype is assigned deterministically based on whether the corresponding number is lower than, higher than, or within the normal range. For example, Figure 1 shows that the lexical candidate "pyrexia" is extracted and then connected with the numeric entity "temperature" by contextualized embeddings. As the number "102°F" is higher than the upper bound of temperature "99.1," the phenotype *Fever (HP:0001945)* is assigned.

Implementation details

First, the syntactic analysis to extract lexical candidates is built by using Stanford Stanza.[29,30] The extraction focuses on nouns, adjectives, and verbs for lexical candidates, which are defined as "NOUN," "PROPN," "ADJ," and "VERB" by the part-of-speech (POS) tagger and the extraction is optimized to capture multiword phrases like "heart rate" by the compound relation. Second, the STS model is built by Sentence Transformers[31] to fine-tune ClinicalBERT up to 4 epochs using the default hyperparameters (Accessed in April 2022: https://www.sbert.net/docs/training/overview.html) with batch size 16 and evaluation step 1000. We use mean pooling to obtain embeddings of multi-word synonyms. The threshold for embedding similarity is set as 0.9 empirically. Some other third-party libraries including PyTorch,[32] Pandas,[33,34] and spaCy are also used in the implementation.

Baselines and evaluation methods

The proposed NR model is compared with previous SOTA phenotyping methods. In the unsupervised setting, we consider unsupervised baseline methods including NCBO,[6] NCR,[8] and the unsupervised model by Zhang *et al.*[12] In the supervised setting, we consider ClinicalBERT[27] (which is fine-tuned in separate for phenotyping) and the supervised model by Zhang *et al.*[12] The NCBO, NCR, and fine-tuned ClinicalBERT are selected as they show overall better performance than other baseline phenotyping methods in corresponding settings, as demonstrated by Zhang *et al.*[12]

We decide not to use recent NR models,[15–18] as baseline methods because none of them considers clinical knowledge and it is costly to adapt them to the clinical domain.

For intrinsic evaluation to assess the accuracy of extracting phenotypes, we compare the proposed NR model against the baseline phenotyping methods by micro-averaged Precision, Recall, and F1-score at the document level.

Table 3. Test set statistics (counts) for the unsupervised and supervised setting.

Test set (unsupervised setting)			Test set (supervised setting)		
EHRs	All phenotypes	NR-specific phenotypes	EHRs	All phenotypes	NR-specific phenotypes
705	20926	1121	170	5047	322

EHR: electronic health record; NR: numerical reasoning.
The unsupervised setting test set refers to all the manually annotated EHRs. On the contrary, supervised setting test set is a subset of the unsupervised setting test set as some of the EHRs from the latter are utilized to fine-tune the baseline models. For the intrinsic evaluation, note that we use only the NR-specific phenotypes, as the other phenotypes typically do not have relations with numbers in clinical narratives.

Table 4. Quantitative evaluation in unsupervised setting.

Model	Exact			Generalized		
	Precision	Recall	F1	Precision	Recall	F1
NCBO/NCR/(unsupervised)[12]	0	0	0	0	0	0
NR	0.5176	0.6879	0.5907	0.6479	0.7907	0.7122

NCBO: national center for bomedical ontology; NCR: neural concept recognizer; NR: numerical reasoning.
In the unsupervised setting, NR model significantly outperforms the baselines NCBO, NCR, and (unsupervised).[12] However, interestingly but not surprisingly, the baselines have zero accuracy as they do not reason with numbers in the clinical narratives.

We follow[35] and compute the metrics by two strategies. (1) Exact Matches: only the exact same HPO annotations against the gold standard annotations are considered as correct; (2) Generalized Matches: the gold standard annotations as well as predicted HPO annotations are extended to include all ancestor HPO concepts until the root concept *Phenotypic Abnormality* (*HP:0000118*) (exclusive) in the HPO hierarchy. All the extended HPO annotations are de-duplicated and added to the list of gold standard and predicted HPO annotations, respectively, for evaluation. By generalized matches, the prediction of HPO concepts which are children, neighbors, or ancestors of the target HPO concepts will also receive credits. For extrinsic evaluation on patient stratification, we report Area Under the Receiver Operating Characteristic curve (AUC-ROC), Sensitivity, and Specificity of the proposed NR model for particular diseases, namely, HIV and myocardial infarction (heart attack), based on the extracted phenotypes of the patients.

Before the deployment of the proposed method in the actual clinical setting, the method is subject to systematic debugging, extensive simulation, testing, and validation under the supervision of expert clinicians following related regulatory guidelines.

Results and discussion

This section introduces the dataset for intrinsic evaluation and demonstrates quantitative analysis, qualitative analysis, and ablation studies against baseline methods. The section then discusses the contributions of the extracted phenotypes to stratify patients and discusses using the extracted phenotypes from clinical text to impute missing values in structured data.

Datasets

Our work is based on the clinical textual notes from the MIMIC-III database which is publicly available.[25] For the unsupervised setting, we use 705 textual notes, which have 20,926 gold phenotype annotations as mentioned in Table 3. Three expert clinicians created the gold phenotype annotations with consensus focusing on contextual synonyms of phenotypes such as "high temperature" and "temperature of 102°F" for *Fever (HP:0001945)*. We narrow down these phenotypes to the ones which need NR using the two conditions: (1) phenotypes belong to the list of HPO IDs in Table 2, which require NR and (2) there is a number in the textual spans of the phenotypes. We identify 1121 such phenotype annotations, which we refer to as NR-specific phenotypes. The proposed NR model is compared with the previous unsupervised baseline methods using the test set created in the unsupervised setting.

For the supervised setting, we fine-tune the baseline methods (like ClinicalBERT) using randomly selected 535 out of 705 manually annotated EHRs. Then, we use the remaining 170 EHRs to compare the NR model with these supervised baselines. So, the test set created for the supervised setting is a subset of the test set created for the unsupervised setting. Note that our proposed NR model is strictly unsupervised in nature, but still we rigorously validate its performance by comparing it with the supervised baseline methods.

The datasets for extrinsic evaluation are also created based on MIMIC-III database. The research has been carried out in accordance with the relevant guidelines and regulations for the MIMIC-III data.

Quantitative analysis

For the unsupervised setting, our quantitative results are reported in Table 4, where we compare the NR model with three baselines: NCBO, NCR, and the unsupervised model by Zhang *et al.*[12] As the baselines are not designed to reason with numbers, they have poor performance on the unsupervised test set, with all of them getting straight 0 on all the metrics. The NR model significantly outperforms all the baselines with 69% Recall and 59% F1 scores using exact

Table 5. Quantitative evaluation in supervised setting.

Model	Exact			Generalized		
	Precision	Recall	F1	Precision	Recall	F1
Fine-tuned ClinicalBERT	**0.8235**	0.181	0.2968	**1.000**	0.2229	0.3646
(Supervised)[12]	0.6791	0.6293	0.6532	0.8245	0.7762	0.7996
NR	0.5952	0.7543	0.6654	0.7290	0.8339	0.7780
(Supervised)[12] + NR	0.5921	**0.8448**	**0.6963**	0.7175	**0.9201**	**0.8062**

NR: numerical reasoning; BERT: Bidirectional Encoder Representations from Transformers.
Comparison of our NR model with the supervised baselines in the supervised setting: It shows that the NR model significantly improves recall by extracting more phenotypes even without explicit supervision. Please note that we refer to a subset of unsupervised setting test set as supervised setting, which we create to compare the unsupervised NR model against the supervised baselines.

metrics, while 79% Recall and 71% F1 scores using generalized metrics. We focus on Recall over Precision so as to extract more phenotypes motivated by the downstream patient stratification system. Extracting more phenotypes help in identifying patients missed by the baselines. Clearly, the NR model demonstrates the superior performance in the unsupervised setting without any costly annotated data.

We further validate the unsupervised NR model by comparing it with the previous SOTA supervised baseline methods. The unsupervised NR model is compared with the supervised model by Zhang *et al.*[12] which was built by fine-tuning with annotated data. Table 5 shows the comparison on the supervised test set. Although the supervised model by Zhang *et al.*[12] performs better than its unsupervised variant, our unsupervised NR model still outperforms the supervised baseline with gains of 12.5% and 5.7% on exact and generalized recall, respectively. However, we observe comparable F1 scores due to precision loss, which is an outcome of our preference toward Recall. Moreover, both the models combined reach the best performance improving score by 21.5% and 14.3% on exact and generalized recall, respectively, and 4.3% and 0.7% gains on exact and generalized F1 scores, respectively. Furthermore, we compare the NR model with the fine-tuned ClinicalBERT[27] whose objective is to detect phenotypes. Again, the NR model and supervised model by Zhang *et al.*[12] combined outperforms this baseline, improving scores by 66.4% and 69.7% on exact and generalized recall, respectively and 40% and 44.2% gains on exact and generalized F1 scores, respectively, as mentioned in Table 5. Overall, it demonstrates the impact of the NR model which performs significantly better than the supervised models, which eliminates the requirement of human annotation of phenotypes which is costly, time-consuming, error-prone, and requires huge manual efforts.

Qualitative analysis

We also evaluate the example sentences having a variety of contexts to compare the NR capabilities of the proposed NR model against the baseline methods. In the sentence "patient has a temperature of 103°F," the unsupervised baselines – NCR, NCBO, and (unsupervised)[12] – fail to detect any phenotype. But with the addition of the word *high*, that is, "patient has a high temperature of 102°F," (unsupervised)[12] is able to detect the phenotype *Fever* (*HP:0001945*) correctly but with the incomplete textual span "high temperature" completely ignoring the number *102°F*. It clearly conveys that it relies solely on the context without reasoning with the number, while NCR and NCBO still miss the phenotype. When we simplify the sentence by replacing the word "temperature" with "fever," that is, "patient has a high *fever* of 102°F," all the three unsupervised baseline methods now correctly determine the phenotype *Fever (HP:0001945)*. However, the number is still missing from the predicted textual span. Overall, we conclude that all the unsupervised baselines just rely on the textual content of the clinical narratives and completely disregard the numbers.

In contrast, our NR model accurately extracts the phenotypes along with correct textual spans including numbers for all the three variants of the original sentence. It correctly identifies the target textual spans, that is, "temperature of 102°F," "temperature of 102°F," and "fever of 102°F," for the phenotype *Fever* (*HP:0001945*) for the three sentences above, respectively. Similarly, it correctly detects the textual spans from the sentence "patient has a breathing rate of 27" with the phenotype *Tachypnea* (*HP:0002789*) as well as "patient has a serum creatinine of 1.7" with the phenotype *Elevated serum creatinine* (*HP:0003259*), while the baselines fail to detect the phenotypes as they do not reason with numbers. Overall, the results indicate that the NR model accurately reasons with the numbers in a variety of contexts without needing any explicit supervision. Note that this efficiently addresses one of the key challenges in NR of handling different contexts as we have identified earlier.

In addition, the (supervised) model[12] achieves reasonable accuracy compared with the unsupervised baselines which get straight 0 scores. However, it still lacks the NR capabilities. For example, given the sentence "patient has a temperature of 102°F," even though it correctly predicts the phenotype *Fever* (*HP:0001945*) but if the number is changed from 102°F to 91°F which changes the target phenotype to *Hypothermia* (*HP:0002045*), the (supervised) model[12] still predicts the phenotype as *Fever* incorrectly. Similarly, we observe wrong predictions when we change the target phenotype from *Tachypnea* (*HP:0002789*) to *Bradypnea* (*HP:0046507*) and from *Elevated serum creatinine* (*HP:0003259*) to *Decreased serum creatinine (HP:0012101)*. We analyzed the phenotype frequencies in the MIMIC data using the NR model predictions and found that there is a strong imbalance in the frequency of the phenotypes corresponding to a numeric entity. For example, we find 8387 cases with the phenotype *Fever*,

Table 6. Ablation studies on the unsupervised test set.

NR model with	Exact			Generalized		
	Precision	Recall	F1	Precision	Recall	F1
Keyword-based shallow matching	**0.6854**	0.2641	0.3813	**0.7745**	0.3449	0.4773
Pretrained contextualized embeddings	0.5065	0.3758	0.4314	0.6006	0.465	0.5241
Fine-tuned contextualized embeddings (used by the final NR model)	0.5176	**0.6879**	**0.5907**	0.6479	**0.7907**	**0.7122**

NR: numerical reasoning.
We compare the NR model variants (1) keyword-based shallow matching of lexical candidates with numeric entities, (2) pretrained contextualized embeddings, and (3) fine-tuned contextualized embeddings. The fine-tuned contextualized embeddings perform much better than the other two variants, thus is used in the final NR model.

which is far more common than the phenotype *Hypothermia* with 5275 cases in the data. Thus, we conclude that the model is fine-tuned with a strong bias toward the highly frequent phenotypes. Therefore, scores in Table 5 are inflated for (supervised)[12] due to overestimation of its NR capabilities. This analysis concludes that the supervision itself is not sufficient for a model without additional tailored learning objectives to achieve the NR capabilities.

Finally, our model also addresses the other two challenges we identified earlier. First, our syntactic analysis handles the *accumulation of multiple numeric examples in a condensed context*. For example, given the sentence, "temperature 99.5, blood pressure 140/90, pulse 85," the NR model correctly identifies *Low-grade fever* (*HP:0011134*), *Hypertension* (*HP:0000822*), and *Tachycardia* (*HP:0001649*). Second, as *not all the numbers in clinical text relate to phenotypes*, our syntactic analysis and contextualized embeddings can discard such cases. For example, in the sentence "patient required 4 days of hospitalization," the NR model finds that *4* is not connected to any numeric entity, so no phenotype is predicted.

Finally, we observe some cases where the NR model does not detect the phenotype accurately. For example, given the text sample, "Pt still with scant bibasilar crackles. Sat @ 97% on 2L NG, continuing with oral HTN meds and Dig.," the model detects the phenotype *Abnormal blood oxygen level (HP:0500165)* (used for negation) from the textual span "Sat @ 97%" as 97% falls within the normal reference range for blood oxygen, that is 95–100%. But the correct phenotype is *Hypoxemia* (*HP:0012418*) as the patient temporarily reaches the normal range due to external oxygen supply, which is evident from the text "2L NG."

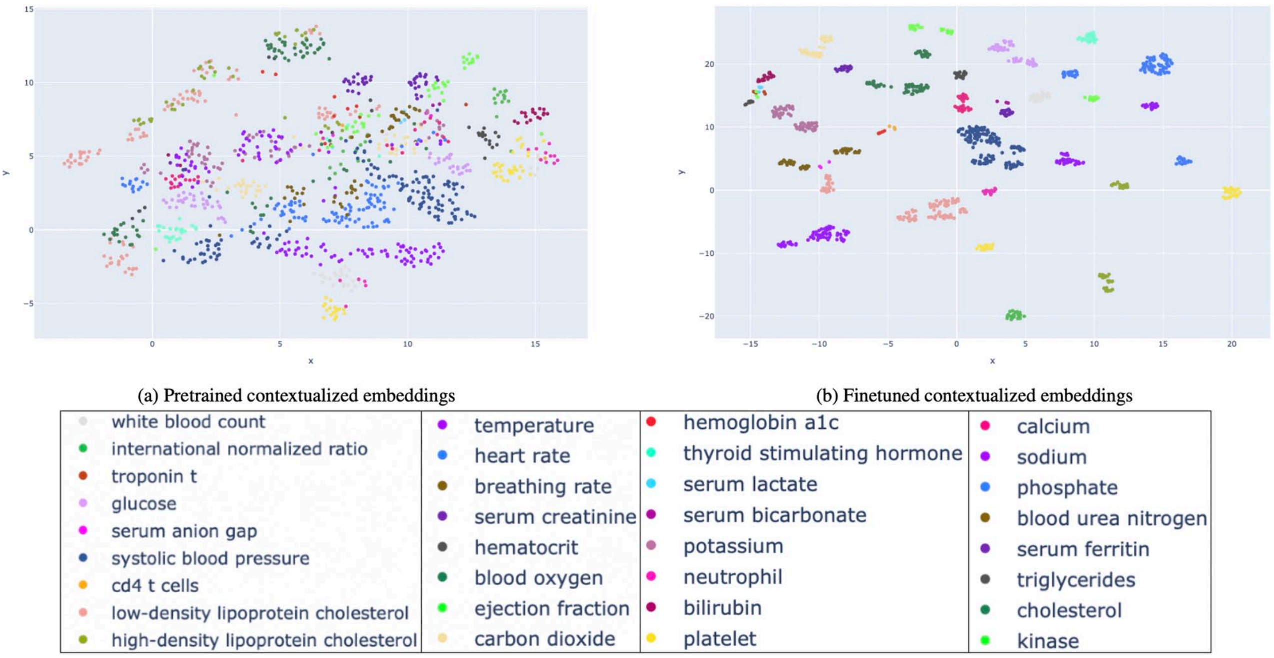
Ablation studies

In order to probe the benefit of contextualized embeddings and the learning objective for fine-tuning in equation (1), we conduct two ablation studies. We assess the need for contextualized embeddings with cosine similarity that is used to connect lexical candidates with numeric entities. Rather, we connect lexical candidates with numeric entities using keyword-based shallow matching. This drops the performance substantially, as reported in Table 6, where the exact Recall drops from 68.8% to 26.4% and F1 drops from 59.1% to 38.1% on the unsupervised test set. Therefore, we need contextualized embeddings to capture the semantics of lexical candidates (corresponding to numeric entities), which may be written in a variety of contexts.

Then, we compare the pretrained contextualized embeddings with the fine-tuned contextualized embeddings as reported in Table 6. As discussed before, the pretrained embeddings are created using the pretrained ClinicalBERT model without any fine-tuning, while the fine-tuned ones are created after fine-tuning the ClinicalBERT using STS equation (1). It is evident from Table 6 that the pretrained embeddings have poor performance with a drop on exact Recall from 68.8% to 37.6% and F1 from 59.1% to 43.1% on the unsupervised test set. We can interpret these results better with Figure 3, which shows the visualization of the pretrained and fine-tuned embeddings of numeric entities and their corresponding UMLS synonyms using uniform manifold approximation and projection (UMAP) dimensionality reduction method.[36] Most of the numeric entities are scattered unevenly in the semantic space created using the pretrained embeddings. For example, there is no clear segregation between the data points for (general) cholesterol, low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol. Contrary, the fine-tuned embeddings form well-segregated data clusters which makes it much easier to predict a corresponding numeric entity of lexical candidates (connected with a number) using cosine similarity without collisions. Overall, the analysis confirms that the proposed learning objective for fine-tuning in equation (1) is needed to connect lexical candidates with numeric entities effectively, as the pretrained contextualized embeddings are not sufficient.

Contribution of NR phenotypes to patient stratification systems

We validate the impact of the NR model with patient stratification systems where we predict whether patients are diagnosed or at risk of particular diseases, namely, HIV and myocardial infarction (heart attack). We use the MIMIC-III data and extract these patients using ICD codes, as mentioned in Table 7. We divide the data (number of admissions) into train and test splits as per Table 8. Later, we use the train set to train a random forest classifier[37] for each disease with the phenotypes extracted from all the MIMIC-III discharge summaries using the best model from Table 5, that is, (supervised)[12] + NR. We also compare our model with a couple of previous SOTA baselines – NCR and fine-tuned ClinicalBERT. The classifiers are evaluated on the test set using the metrics – AUC-ROC, Sensitivity, and Specificity as shown in Table 9.



(a) Pretrained contextualized embeddings

(b) Finetuned contextualized embeddings

(c) Color codes for numeric entities

Figure 3. Visualizing pretrained and fine-tuned contextualized embeddings of numeric entities along with their UMLS synonyms obtained by pretrained and fine-tuned ClinicalBERT, respectively, using UMAP plots. We observe that fine-tuned embeddings result in better differentiation of numeric entities in the semantic space. Our NR model exploits this differentiation to identify the phenotypes more accurately.

Table 7. Extrinsic evaluation: diseases statistics using MIMIC-III database.

Disease	ICD-9 codes	Number of admissions	
		Positive	Negative
HIV	042 and 079.53	538	58438
Myocardial infarction (heart attack)	All subcodes of 410	5430	53546

ICD: International Classification of Diseases; MIMIC: medical information mart for intensive care.
Positive admissions are extracted from MIMIC-III database using ICD-9 codes, while the rest of the admissions are marked as negative.

Table 8. Extrinsic evaluation: train and test split for patient stratification systems.

Disease	Train			Test		
	All	Positive	Negative	All	Positive	Negative
HIV	29,637	307	29,330	29,339	231	29,108
Myocardial infarction (heart attack)	29,637	2767	26,870	29,339	2663	26,676

Number of admissions divided into train and test split for patient stratification systems.

Table 9. Extrinsic evaluation: quantitative evaluation for patient stratification systems.

Disease	Model	AUC-ROC	Sensitivity	Specificity
HIV	NCR	0.701	0.415	**0.986**
	Fine-tuned ClinicalBERT	0.901	0.818	0.983
	(Supervised)^{12} + NR	**0.966**	**0.952**	0.980
Myocardial infarction (heart attack)	NCR	0.831	0.822	0.840
	Fine-tuned ClinicalBERT	0.839	0.862	0.817
	(Supervised)^{12} + NR	**0.870**	**0.888**	**0.853**

AUC-ROC: area under the receiver operating characteristic curve; NCR: neural concept recognizer; NR: numerical reasoning; BERT: Bidirectional Encoder Representations from Transformers.
Evaluating the benefit of the proposed NR model on the downstream use case to stratify patients at risk of particular diseases using the best model from Table 5, that is, (supervised)^{12} + NR comparing against the baselines – NCR and Fine-tuned ClinicalBERT.

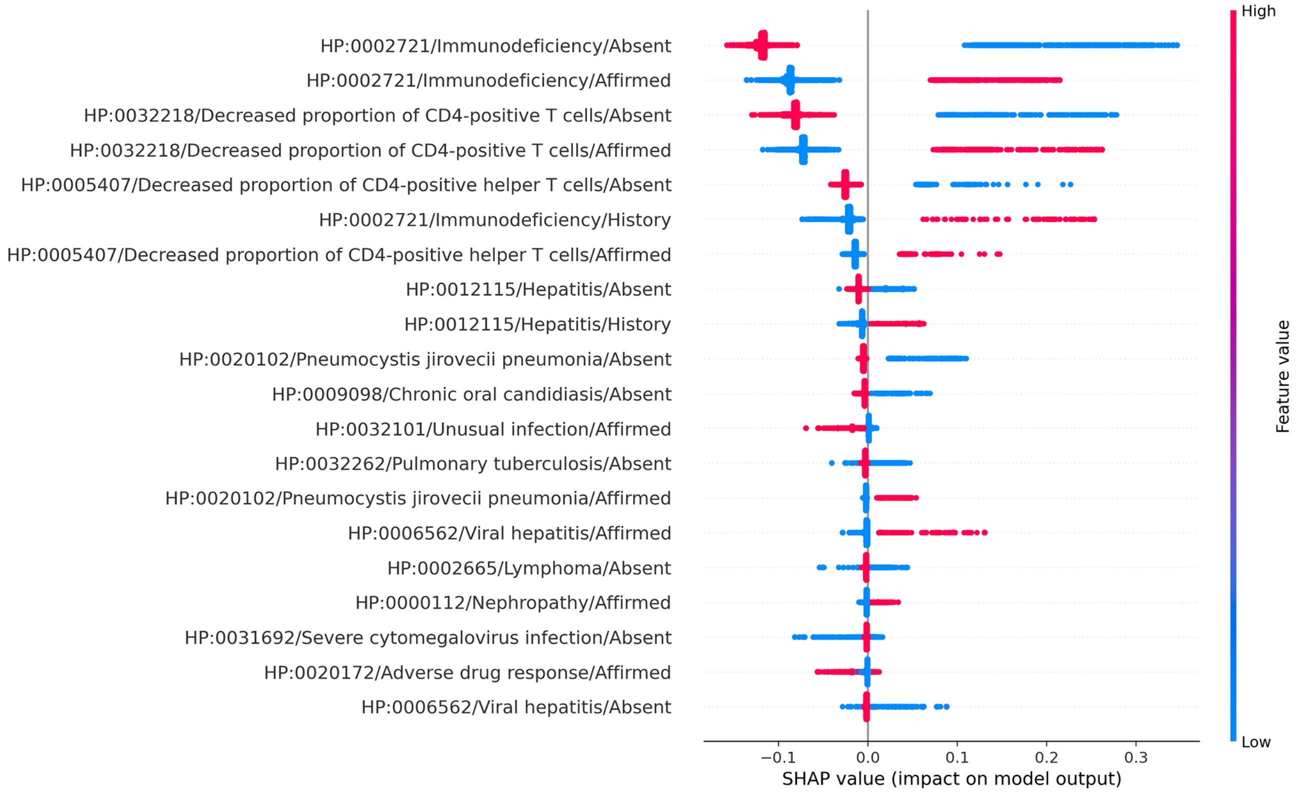


Figure 4. HIV patient stratification system: Technical evaluation using Shapley scores using top phenotypes sorted in the descending order of importance. Here, the red color indicates the presence of a phenotype, while blue indicates the absence. On the SHAP axis, dots on the right-hand side refer to patients with high probability of HIV, while the left-hand side dots refer to low probability. Density of dots indicate the number of patients. This clearly demonstrates that the phenotype "HP:0032218/Decreased proportion of CD4-positive T cells," which requires numerical reasoning, is among the top indicators to detect a patient with HIV.

For HIV, our model significantly outperforms all the baselines on AUC-ROC and Sensitivity metrics by obtaining the scores of 96.6% and 95.2%, respectively. All the models have similar Specificity. In the case of myocardial infarction, our model outperforms all the baselines on all the metrics by obtaining an AUC-ROC of 87.0%, Sensitivity of 88.8%, and Specificity of 85.3%. Overall, it demonstrates the capability of our model to precisely extract the positive patients from a large pool of candidates.

In order to analyze the impact of the NR model on the prediction performance, we analyze the top phenotype features that the classifiers rely on to identify the patients. We use Shapley (SHAP)[38] analysis as shown in Figures 4 and 5 for HIV and myocardial infarction, respectively. We validated the clinical importance of these phenotypes for each disease using the clinician scores (refer Table 10) as shown in Tables 11 and 12 for HIV and myocardial infarction, respectively. These scores are assigned to the phenotypes by three expert clinicians where a score of 1 means that a phenotype is irrelevant to a disease, while a score of 5 means that the phenotype is the most important clinical indicator for a disease.

In the case of HIV, the phenotype "HP:0032218/Decreased proportion of CD4-positive T cells" extracted by NR model is among the top features used by the patient stratification system as is evident from its SHAP plot. This phenotype is further verified by the clinicians who marked it as one of the most important clinical indicators for HIV with a clinician score of 5. Similarly, for myocardial infarction, the phenotypes "HP:0410174/Increased troponin T level in blood" and "HP:0012666/Severely reduced ejection fraction" extracted by NR model are among the top features which are further verified by the clinicians marking them as one of the most important clinical indicators for myocardial infarction using the clinician scores of 5 and 4, respectively. We hypothesize that the phenotypes captured by the NR model contain important bedside measurements written in clinical textual notes, which help the classifiers for better patient stratification.

Imputation of missing structured data using NR phenotypes

We conducted an analysis to impute the missing structured data from the MIMIC-III dataset with the numeric values present in the phenotypes extracted from the textual notes by the NR model. We aggregate all the structured data recorded at different timesteps for an intensive care unit (ICU) admission and mark all the empty or negative aggregated values as missing data. Then, we extract the phenotypes from the corresponding discharge summaries and impute the missing data with the numeric values from the phenotypic text. For example, if temperature is missing from the structured data for an admission and if we detect a phenotype "Fever" in its discharge summary with the phenotypic text "temperature of 99 F," then we impute the missing value with the number "99."

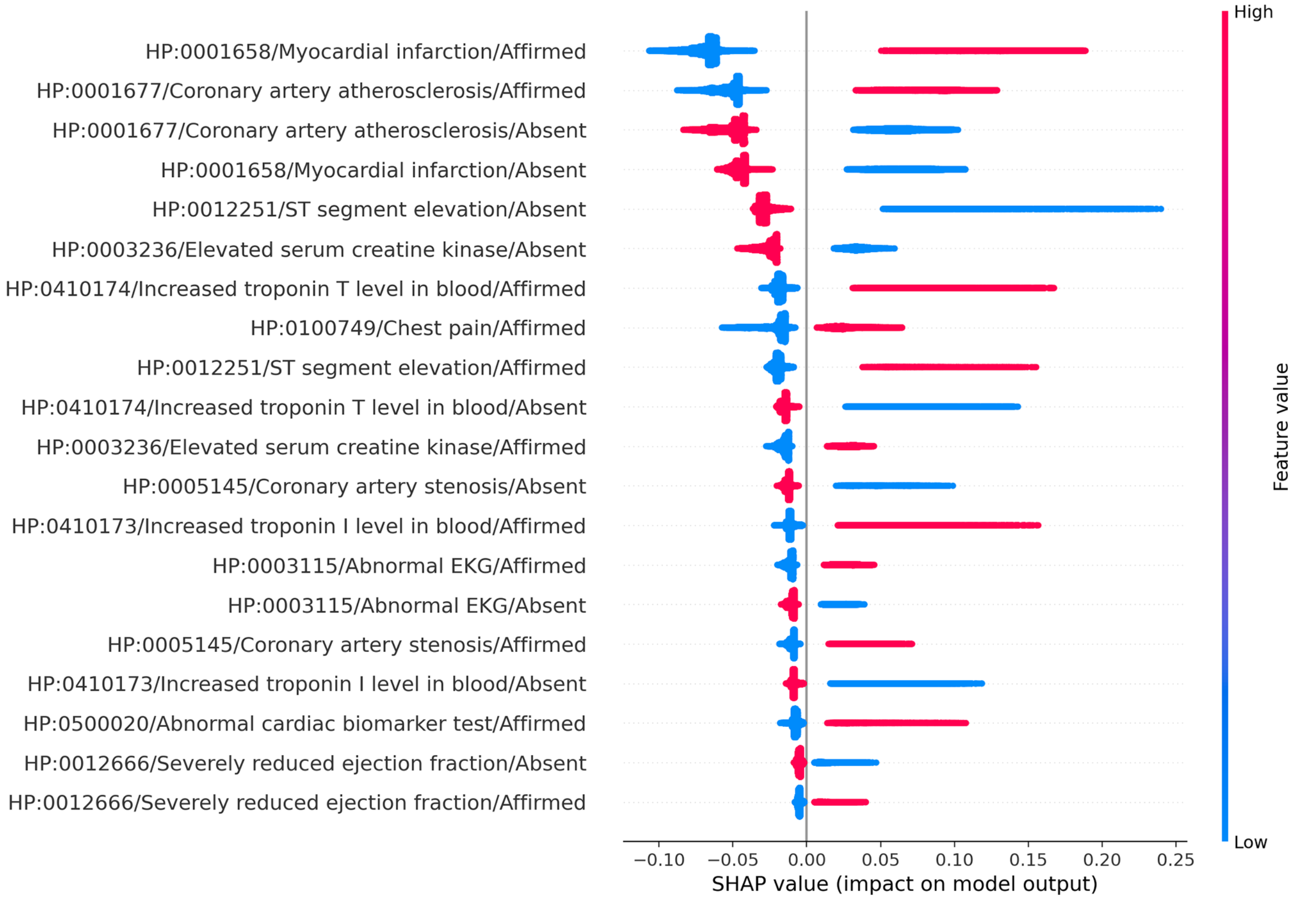


Figure 5. Myocardial infarction (heart attack) patient stratification system: Technical evaluation using Shapley scores using top phenotypes sorted in the descending order of importance. Here, the red color indicates the presence of a phenotype, while blue indicates the absence. On the SHAP axis, the right-hand side dots refer to patients with high probability of myocardial infarction, while the left-hand side dots refer to low probability. Density of dots indicate the number of patients. This clearly demonstrates that the phenotypes “HP:0410174/Increased troponin T level in blood” and “HP:0012666/Severely reduced ejection fraction,” which require numerical reasoning, are among the top indicators to detect a patient with myocardial infarction.

Table 10. Extrinsic evaluation: clinician score to validate patient stratification systems.

Clinicians’ score	Phenotype importance
1	Irrelevant
2	Low
3	Moderate
4	High
5	Very high

Clinicians’ scores to validate phenotype features for patient stratification systems: A score of 1 means that a phenotype is irrelevant to a disease, while a score of 5 means that the phenotype is the most important clinical indicator for a disease.

We impute the structured data for several features including breathing rate, calcium, hematocrit, and so on, as shown in Table 13 where the imputation can fill at least 25% of the missing values. It drastically improves the quality of the structured data, as the imputed values are the actual measurements instead of estimated averages or median values of all the available values. Structured data is useful for several downstream biomedical applications such as ICU in-hospital mortality and length of stay prediction,[39] which can benefit the most with the improved data quality.

Generalizability

We have shown the significance of the proposed NR model for the phenotype annotation, and it can be potentially generalized to other healthcare or general NLP tasks. For example, in the biomedical domain, the model can be adopted to extract drug doses from clinical text by (1) extracting numeric dosage and drug names, (2) matching drug names with entities in a predefined external drug database by contextualized embeddings, and (3) reasoning what the dosage indicates. Similarly, the model can also be applied in other NLP tasks such as summarization (with numbers in documents and summaries) and question answering (understanding numbers in questions, context, and answers).

Limitations

The major limitation is the discrepancy due to the boundary cases. For example, 99.1°F is the upper normal temperature limit, so strictly speaking 99.2°F is Fever. However, it is not always true in real-world practice as sometimes the lower and upper bounds are defined vaguely and may vary place to place. The lower and upper bounds also vary with age and gender, which means the external knowledge may be dynamic. Therefore, in the future we plan to address those

Table 11. Extrinsic evaluation: clinical validation for HIV patient stratification system.

Phenotype	Clinician Score (1–5)	Is a comorbidity?
HP:0002721/Immunodeficiency	5	No
HP:0032218/Decreased proportion of CD4-positive T cells	5	No
HP:0005407/Decreased proportion of CD4-positive helper T cells	5	No
HP:0012115/Hepatitis	3	Yes
HP:0020102/Pneumocystis jirovecii pneumonia	5	Yes
HP:0009098/Chronic oral candidiasis	5	Yes
HP:0032101/Unusual infection	3	No
HP:0032262/Pulmonary tuberculosis	4	Yes
HP:0006562/Viral hepatitis	3	Yes
HP:0002665/Lymphoma	3	Yes
HP:0000112/Nephropathy	4	No
HP:0031692/Severe cytomegalovirus infection	5	Yes
HP:0020172/Adverse drug response	1	No

Clinical evaluation by three expert clinicians using Clinician scores (1–5) (refer Table 10) are used to mark the importance of the top phenotypes as clinical indicators for the HIV. Comorbidity status of these phenotypes with respect to HIV is also marked. All the phenotypes extracted by numerical reasoning model are bold. This clearly demonstrates that the phenotype "HP:0032218/Decreased proportion of CD4-positive T cells" which requires numerical reasoning is among the top indicators to detect a patient with HIV, which is further verified by clinicians marking it as one of the most important clinical indicators for HIV.

Table 12. Extrinsic evaluation: clinical validation for myocardial infarction (heart attack) patient stratification system.

Phenotype	Clinician score (1–5)	Is a comorbidity?
HP:0001658/myocardial infarction	5	Yes
HP:0001677/Coronary artery atherosclerosis	3	Yes
HP:0003236/Elevated serum creatine kinase	1	No
HP:0012251/ST segment elevation	5	Yes
HP:0005145/Coronary artery stenosis	5	Yes
HP:0410174/Increased troponin T level in blood	5	Yes
HP:0003115/Abnormal EKG	4	Yes
HP:0410173/Increased troponin I level in blood	5	Yes
HP:0100749/Chest pain	3	Yes
HP:0500020/Abnormal cardiac biomarker test	5	Yes
HP:0012666/Severely reduced ejection fraction	4	Yes

EKG: electrocardiogram.
Clinical evaluation by three expert clinicians using Clinician scores (1–5) (refer Table 10) are used to mark the importance of the top phenotypes as clinical indicators for myocardial infarction. Comorbidity status of these phenotypes with respect to myocardial infarction is also marked. All the phenotypes extracted by numerical reasoning model are given in bold. This clearly demonstrates that the phenotypes "HP:0410174/Increased troponin T level in blood" and "HP:0012666/Severely reduced ejection fraction," which require numerical reasoning are among the top indicators to detect a patient with myocardial infarction, which are further verified by clinicians marking them as one of the most important clinical indicators for myocardial infarction.

Table 13. Extrinsic evaluation: imputation of missing structured data in the MIMIC-III corpus to enrich data quality.

Structured data	Total missing admissions	% of missing admissions imputed by NR phenotypes
Breathing rate	57,947	39.774%
Calcium	53591	38.159%
Carbon dioxide	31,520	26.396%
Heart rate	6677	31.631%
Hematocrit	8454	31.595%
International normalized ratio	22,663	26.479%
Phosphate	43,688	30.331%
Platelet	9869	26.376%
Serum creatinine	58,976	65.001%
Systolic blood pressure	10,400	45.385%
White blood count	16,140	45.273%

MIMIC: medical information mart for intensive care; NR: numerical reasoning.
Imputation of missing structured data in the MIMIC-III corpus having a total of 58,976 admissions using phenotypes extracted by NR model: Imputing missing values using numbers from textual notes is critical to improve the structured data quality, which is useful for several downstream biomedical applications such as ICU in-hospital mortality and length of stay prediction.

Table 14. Numeric entities with normal reference range (complete table).

ID	Numeric entity	Abbreviation	Unit	Normal reference range	
				Lower bound	Upper bound
0	Temperature	Temp	Celsius	36.4	37.3
0	Temperature	Temp	Fahrenheit	97.5	99.1
1	Heart rate	Heart rate	Beats per minute (bpm)	60	80
2	Breathing rate	Breathing rate	Breaths per minute	12	20
3	Serum creatinine	Serum creatinine	mg/dL	0.6	1.2
3	Serum creatinine	Serum creatinine	micromoles/L	53	106.1
4	Hematocrit	Hct	%	41	48
5	Blood oxygen	O_2	%	95	100
6	Ejection fraction	EF	%	50	100
7	Carbon dioxide	CO_2	mEq/L	23	29
8	White blood count	WBC	*1000	4.5	11
8	White blood count	WBC	*1	4500	11,000
9	International normalized ratio	INR	*1	0.8	1.1
10	Troponin T	Trop t	ng/ml	0	0.04
11	Glucose	Glucose	mmol/L	3.9	5.6
11	Glucose	Glucose	mg/dL	70	100
12	Serum anion gap	AG	mEq/L	3	10
13	Systolic blood pressure	SBP	mmHG	90	139
14	CD4T cells	CD4	Cells/mm^3	500	1500
15	Low-density lipoprotein cholesterol	Ldl	mmol/L	0	2.6
15	Low-density lipoprotein cholesterol	Ldl	mg/dL	70	100
16	High-density lipoprotein cholesterol	Hdl	mmol/L	1.5	5
16	High-density lipoprotein cholesterol	Hdl	mg/dL	60	1000
17	Hemoglobin a1c	Hba1c	%	0	5.6
18	Thyroid stimulating hormone	Tsh	mIU/L	0.5	5
19	Serum lactate	Lactate	mmol/L	0.5	2.2
20	Serum bicarbonate	Hco3	mEq/L	23	30
21	Potassium	k	mmol/L	3.5	5.5
22	Neutrophil	Anc	10^9/L	2	7.5
22	Neutrophil	Anc	Cells per microliter	2000	7500
23	Bilirubin	Bilirubin	mg/dL	0.3	1.2
24	Platelet	Platelet	*1000	150	450
24	Platelet	Platelet	*1	150,000	450,000
25	Calcium	Ca	mmol/L	2	2.5
25	Calcium	Ca	mg/dL	8	10
26	Sodium	Na	mmol/L	136	145
27	Phosphate	Phosphate	mmol/L	0.97	1.45
27	Phosphate	Phosphate	mg/dL	3	4.5
28	Blood urea nitrogen	Bun	mmol/L	2.5	7.1
28	Blood urea nitrogen	Bun	mg/dL	7	20
29	Serum ferritin	Serum ferritin	ng/mL	20	150
30	Triglycerides	Tg	mmol/L	0	1.7
30	Triglycerides	Tg	mg/dL	50	150
31	Cholesterol	Chol	mmol/L	0	5
31	Cholesterol	Chol	mg/dL	50	200
32	Kinase	Ck	U/L	25	200

A complete table of numeric entities that are used in the study with normal reference range and units. The ID column corresponds to that in Table 15.

cases by fine-tuning a dynamic precision level to optimize prediction performance. Other than the boundary cases, we have restricted our method to work at the sentence-level context. Extending to the document level and contextual reasoning on longer context along with NR will further improve the model. For instance, if the temperature is recorded as 98°F (normal temperature) at the beginning of a clinical note, but the note mentions toward the end that it increases by 3°F, then ideally a model should be able to capture fever by document level reasoning as the temperature 101°F means fever. The proposed model also relies on pre-extracted external knowledge that is manually curated. Ways to extract this knowledge automatically could be explored further.

Future works

The proposed model can be potentially generalized to other biomedical NLP tasks that require NR from text. The model can be further extended to consider document-level context and dynamic external knowledge base.

Table 15. Numeric entities and corresponding phenotypes (complete table).

ID	Numeric entity	Number lower than the lower bound (affirmed)		Number higher than the upper bound (affirmed)		Number inside normal range (negated)	
		HPO ID	HPO Name	HPO ID	HPO Name	HPO ID	HPO Name
0	Temperature	HP:0002045	Hypothermia	HP:0001945	Fever	HP:0004370	Abnormality of temperature regulation
1	Heart rate	HP:0001662	Bradycardia	HP:0001649	Tachycardia	HP:0011675	Arrhythmia
2	Breathing rate	HP:0046507	Bradypnea	HP:0002789	Tachypnea	HP:0002793	Abnormal pattern of respiration
3	Serum creatinine	HP:0012101	Decreased serum creatinine	HP:0003259	Elevated serum creatinine	HP:0012100	Abnormal circulating creatinine concentration
4	Hematocrit	HP:0031851	Reduced hematocrit	HP:0001899	Increased hematocrit	HP:0031850	Abnormal hematocrit
5	Blood oxygen	HP:0012418	Hypoxemia	HP:0012419	Hyperoxemia	HP:0500165	Abnormal blood oxygen level
6	Ejection fraction	HP:0012664	Reduced ejection fraction			HP:0012664	Reduced ejection fraction
7	Carbon dioxide	HP:0012417	Hypocapnia	HP:0012416	Hypercapnia	HP:0500164	Abnormal blood carbon dioxide level
8	White blood count	HP:0001882	Leukopenia	HP:0001974	Leukocytosis	HP:0011893	Abnormal leukocyte count
9	International normalized ratio	HP:0032198	Decreased prothrombin time	HP:0008151	Prolonged prothrombin time	HP:0032199	Abnormal prothrombin time
10	Troponin T			HP:0410174	Increased troponin T level in blood	HP:0410174	Increased troponin T level in blood
11	Glucose	HP:0001943	Hypoglycemia	HP:0003074	Hyperglycemia	HP:0011015	Abnormal blood glucose concentration
12	Serum anion gap	HP:0031963	Decreased serum anion gap	HP:0031962	Elevated serum anion gap	HP:0031961	Abnormal serum anion gap
13	Systolic blood pressure	HP:0002615	Hypotension	HP:0000822	Hypertension	HP:0030972	Abnormal systemic blood pressure
14	CD4T cells	HP:0032218	Decreased proportion of CD4-positive T cells	HP:0032219	Increased proportion of CD4-positive T cells	HP:0031392	Abnormal proportion of CD4-positive T cells
15	Low-density lipoprotein cholesterol	HP:0003563	Decreased LDL cholesterol concentration	HP:0003141	Increased LDL cholesterol concentration	HP:0003141	Increased LDL cholesterol concentration
16	High-density lipoprotein cholesterol	HP:0003233	Decreased HDL cholesterol concentration	HP:0012184	Increased HDL cholesterol concentration	HP:0031888	Abnormal HDL cholesterol concentration
17	Hemoglobin A1c			HP:0040217	Elevated hemoglobin A1c	HP:0040217	Elevated hemoglobin A1c
18	Thyroid stimulating hormone	HP:0000836	Hyperthyroidism	HP:0000821	Hypothyroidism	HP:0031508	Abnormal thyroid hormone level
19	Serum lactate			HP:0002151	Increased serum lactate	HP:0002151	Increased serum lactate
20	Serum bicarbonate	HP:0032066	Decreased serum bicarbonate concentration	HP:0032067	Elevated serum bicarbonate concentration	HP:0032065	Abnormal serum bicarbonate concentration
21	Potassium	HP:0002900	Hypokalemia	HP:0002153	Hyperkalemia	HP:0011042	Abnormal blood potassium concentration
22	Neutrophil	HP:0001875	Neutropenia	HP:0011897	Neutrophilia	HP:0011991	Abnormal neutrophil count
23	Bilirubin	HP:0033480	Hypobilirubinemia	HP:0002904	Hyperbilirubinemia	HP:0033479	Abnormal circulating bilirubin concentration
24	Platelet	HP:0001873	Thrombocytopenia	HP:0001894	Thrombocytosis	HP:0011873	Abnormal platelet count
25	Calcium	HP:0002901	Hypocalcemia	HP:0003072	Hypercalcemia	HP:0004363	Abnormal circulating calcium concentration
26	Sodium	HP:0002902	Hyponatremia	HP:0003228	Hypernatremia	HP:0010931	Abnormal blood sodium concentration
27	Phosphate	HP:0002148	Hypophosphatemia	HP:0002905	Hyperphosphatemia	HP:0100529	Abnormal blood phosphate concentration
28	Blood urea nitrogen	HP:0031969	Reduced blood urea nitrogen	HP:0003138	Increased blood urea nitrogen	HP:0031970	Abnormal blood urea nitrogen concentration
29	Serum ferritin	HP:0012343	Decreased circulating ferritin concentration	HP:0003281	Increased circulating ferritin concentration	HP:0040133	Abnormal circulating ferritin concentration
30	Triglycerides	HP:0012153	Hypotriglyceridemia	HP:0002155	Hypertriglyceridemia	HP:0002155	Hypertriglyceridemia
31	Cholesterol	HP:0003146	Hypocholesterolemia	HP:0003124	Hypercholesterolemia	HP:0003107	Abnormal circulating cholesterol concentration
32	Kinase			HP:0003236	Elevated circulating creatine kinase concentration	HP:0003236	Elevated circulating creatine kinase concentration

HDL: high-density lipoprotein; HPO: human phenotype ontology; LDL: low-density lipoprotein.
A complete table of numeric entities that are used in the study with phenotype labels (including HPO ID and HPO name). The ID column corresponds to Table 14.

AUTHORS' CONTRIBUTIONS

All the authors participated in the conceptualization and design of the study. AT, JZ, and JI worked on building the methodology and writing the manuscript. A.T. conducted the experiments, validated the results, and performed the analysis. JZ provided the resources and worked on data curation. JI, VG, and YG reviewed the manuscript. VG and YG supervised the work. AT and JZ contributed equally to this paper.

ACKNOWLEDGEMENTS

The authors thank Dr Garima Gupta, Dr Deepa (MRSH), and Dr Ashok (MS) for helping to create gold standard phenotype annotation data, validating the external knowledge for numerical reasoning, and providing feedback on the phenotype indicators for the diseases.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: AT, JZ, VG, and YG are employed at Pangaea Data Limited. JI is employed at Queen Mary University of London, but this study was conducted when she was employed at Pangaea Data Limited. YG is also employed at Imperial College London and Hong Kong Baptist University, but he supervised this study at Pangaea Data Limited.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is solely funded by Pangaea Data Limited.

DATA AVAILABILITY

The MIMIC-III dataset is publicly available at https://mimic.mit.edu/ (accessed in November 2021). The code, gold labels, and other data cannot be shared publicly due to their proprietary nature. The implementation details are already described in the paper.

ORCID ID

Ashwani Tanwar https://orcid.org/0000-0001-9032-4348

REFERENCES

1. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;**21**:221–30
2. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat* 2012;**33**:777–80
3. World Health Organization. *ICD-10: international statistical classification of diseases and related health problems: tenth revision*. Geneva: World Health Organization, 2004
4. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229–36
5. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668/
6. Jonquet C, Shah N, Youn C, Callendar C, Storey MA, Musen M. NCBO annotator: semantic annotation of biomedical data. In: *International semantic web conference, poster and demo session*, vol. 110, Washington, DC, 25 October 2009
7. Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, Genetti CA, Brownstein CA, Schmitz-Abe K, Schoch K, Cope H, Signer R, Undiagnosed Diseases Network, Martinez-Agosto JA, Shashi V, Beggs AH, Wheeler MT, Bernstein JA, Bejerano G. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med* 2019; **21**:1585–93
8. Arbabi A, Adams DR, Fidler S, Brudno M. Identifying clinical terms in medical text using Ontology-Guided machine learning. *JMIR Med Inform* 2019;**7**:e12596
9. Kraljevic Z, Bean D, Mascio A, Roguski L, Folarin A, Roberts A, Bendayan R, Dobson R. MedCAT – Medical Concept Annotation Tool, 2019, https://arxiv.org/abs/1912.10166
10. Tiwari P, Uprety S, Dehdashti S, Hossain MS. TermInformer: unsupervised term mining and analysis in biomedical literature. *Neural Comput Appl*. Epub ahead of print 16 September 2020. DOI: 10.1007/s00521-020-05335-2
11. Kohler S, Gargano MA, Matentzoglu N, Carmody L, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD, Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnett MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yüksel Z, Helbig I, Mungall CJ, Haendel MA, Robinson PN. The human phenotype ontology in 2021. *Nucleic Acids Res* 2021;**49**:D1207–17
12. Zhang J, Bolanos L, Li T, Tanwar A, Freire G, Yang X, Ive J, Gupta V, Guo Y. Self-supervised detection of contextual synonyms in a multi-class setting: phenotype annotation use case. In: *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP)*, Punta Cana, Dominican Republic, 7–11 November 2021, pp. 8754–69. New York: Association for Computational Linguistics
13. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, NAACL-HLT 2019*, vol. 1 (long and short papers). New York: Association for Computational Linguistics, 2019, pp. 4171–86
14. Zhang Q, Wang L, Yu S, Wang S, Wang Y, Jiang J, Lim E-P. NOAHQA: numerical reasoning with interpretable graph question answering dataset. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, November 2021, pp. 4147–61. New York: Association for Computational Linguistics, https://aclanthology.org/2021.findings-emnlp.350
15. Thawani A, Pujara J, Ilievski F. Numeracy enhances the literacy of language models. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*, Punta Cana, Dominican Republic, 7–11 November 2021, pp. 6960–7 [Online]. New York: Association for Computational Linguistics, https://aclanthology.org/2021.emnlp-main.557
16. Duan H, Yang Y, Tam KY. Learning numeracy: a simple yet effective number embedding approach using knowledge graph. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, November 2021, pp. 2597–602. New York: Association for Computational Linguistics, https://aclanthology.org/2021.findings-emnlp.221
17. Saha A, Joty S, Hoi SC. Weakly supervised neuro-symbolic module networks for numerical reasoning over text. *AAAI* 2022;**36**:11238–47, https://ojs.aaai.org/index.php/AAAI/article/view/21374 (accessed 6 July 2022)
18. Jin Z, Jiang X, Wang X, Liu Q, Wang Y, Ren X, Qu H. NumGPT: improving numeracy ability of generative pre-trained models. *CoRR* 2021:abs/2109.03137, https://arxiv.org/abs/2109.03137
19. Geva M, Gupta A, Berant J. Injecting numerical reasoning skills into language models. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds)

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. New York: Association for Computational Linguistics, 2020, pp. 946–58

20. Dua D, Wang Y, Dasigi P, Stanovsky G, Singh S, Gardner M. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. In: Burstein J, Doran C, Solorio T (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, **vol. 1** (Long and short papers). New York: Association for Computational Linguistics, 2019, pp. 2368–78
21. Hu M, Peng Y, Huang Z, Li D. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In: Inui K, Jiang J, Ng V, Wan X (eds) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. New York: Association for Computational Linguistics, 2019, pp. 1596–606
22. Ran Q, Lin Y, Li P, Zhou J, Liu Z. NumNet: machine reading comprehension with numerical reasoning. In: Inui K, Jiang J, Ng V, Wan X (eds) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. New York: Association for Computational Linguistics, 2019, pp. 2474–84
23. Sushil M, Suster S, Daelemans W. Are we there yet? Exploring clinical domain knowledge of BERT models. In: *Proceedings of the 20th Workshop on Biomedical Language Processing* [Online]. New York: Association for Computational Linguistics, 2021, pp. 41–53, https://aclanthology.org/2021.bionlp-1.5
24. Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJS, DeMare LE, Devereau AD, de Vries BBA, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, Laulederkind SJF, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MWM, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JOB, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. The human phenotype ontology in 2017. *Nucleic Acids Res* 2016;**45**:D865–76
25. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;**3**:1–9
26. Moon S, Pakhomov S, Liu N, Ryan JO, Melton GB. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *J Am Med Inform Assoc* 2014;**21**:299–307
27. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, McDermott M. Publicly available Clinical BERT embeddings. In: *Proceedings of the 2nd clinical natural language processing workshop*, Minneapolis, MN, June 2019, pp. 72–8. New York: Association for Computational Linguistics, https://aclanthology.org/W19-1909
28. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:267–70
29. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a Python natural language processing toolkit for many human languages. In: *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, July 2020, pp. 101–8, https://nlp.stanford.edu/pubs/qi2020stanza.pdf
30. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. *J Am Med Inform Assoc* 2021;**28**:1892–9
31. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X (eds) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. New York: Association for Computational Linguistics, 2019, pp. 3980–90
32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: an imperative style, high-performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8 December 2019, pp. 8026–37. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
33. McKinney W. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in science conference*, 28 June 2010, vol. 445, no. 1, pp. 56–61, https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf
34. McKinney W. pandas: a foundational Python library for data analysis and statistics. *Python High Perform Sci Comput* 2011;**14**:1–9
35. Liu C, Ta CN, Rogers JR, Li Z, Lee J, Butler AM, Shang N, Kury FSP, Wang L, Shen F, Liu H, Ena L, Friedman C, Weng C. Ensembles of natural language processing systems for portable phenotyping solutions. *J Biomed Inform* 2019;**100**:103318, http://www.sciencedirect.com/science/article/pii/S1532046419302370
36. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints* 2018, https://arxiv.org/abs/1802.03426
37. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32, https://link.springer.com/article/10.1023/A:1010933404324
38. Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) *Advances in Neural Information Processing Systems 30: annual conference on Neural Information Processing Systems 2017*. NeurIPS, 2017, pp. 4765–74, https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
39. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019;**6**:1–18

(Received May 11, 2022, Accepted July 12, 2022)