**RESEARCH**

# Ontology-driven and weakly supervised rare disease identification from clinical notes

Hang Dong[1,2,3*], Víctor Suárez-Paniagua[1,2], Huayu Zhang[4], Minhong Wang[5], Arlene Casey[4], Emma Davidson[6], Jiaoyan Chen[7], Beatrice Alex[8], William Whiteley[2,6] and Honghan Wu[2,5*]

## Abstract

**Background**  Computational text phenotyping is the practice of identifying patients with certain disorders and traits from clinical notes. Rare diseases are challenging to be identified due to few cases available for machine learning and the need for data annotation from domain experts.

**Methods**  We propose a method using ontologies and weak supervision, with recent pre-trained contextual representations from Bi-directional Transformers (e.g. BERT). The ontology-driven framework includes two steps: (i) Text-to-UMLS, extracting phenotypes by contextually linking mentions to concepts in Unified Medical Language System (UMLS), with a Named Entity Recognition and Linking (NER+L) tool, SemEHR, and weak supervision with customised rules and contextual mention representation; (ii) UMLS-to-ORDO, matching UMLS concepts to rare diseases in Orphanet Rare Disease Ontology (ORDO). The weakly supervised approach is proposed to learn a phenotype confirmation model to improve Text-to-UMLS linking, without annotated data from domain experts. We evaluated the approach on three clinical datasets, MIMIC-III discharge summaries, MIMIC-III radiology reports, and NHS Tayside brain imaging reports from two institutions in the US and the UK, with annotations.

**Results**  The improvements in the precision were pronounced (by over 30% to 50% absolute score for Text-to-UMLS linking), with almost no loss of recall compared to the existing NER+L tool, SemEHR. Results on radiology reports from MIMIC-III and NHS Tayside were consistent with the discharge summaries. The overall pipeline processing clinical notes can extract rare disease cases, mostly uncaptured in structured data (manually assigned ICD codes).

**Conclusion**  The study provides empirical evidence for the task by applying a weakly supervised NLP pipeline on clinical notes. The proposed weak supervised deep learning approach requires no human annotation except for validation and testing, by leveraging ontologies, NER+L tools, and contextual representations. The study also demonstrates that Natural Language Processing (NLP) can complement traditional ICD-based approaches to better estimate rare diseases in clinical notes. We discuss the usefulness and limitations of the weak supervision approach and propose directions for future studies.

**Keywords**  Clinical notes, Natural language processing, Ontology matching, Phenotyping, Rare diseases, Weak supervision

*Correspondence:
Hang Dong
hang.dong@cs.ox.ac.uk
Honghan Wu
honghan.wu@ucl.ac.uk
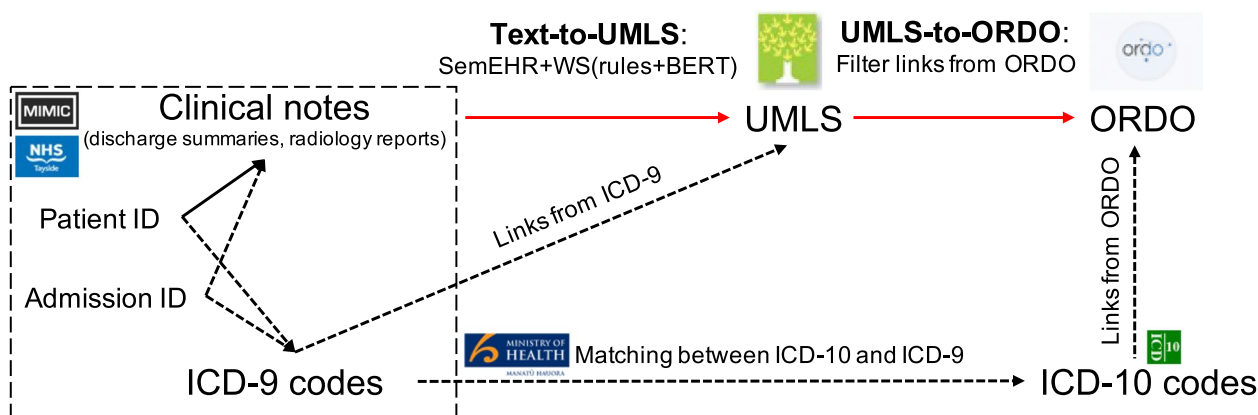Full list of author information is available at the end of the article

**Fig. 1** A pipeline for rare disease identification from clinical notes with ontologies and weak supervision. The upper horizontal lines (in red) show the proposed pipeline based on clinical notes (e.g. discharge summaries and radiology reports in US MIMIC-III and UK NHS Tayside) and ontologies, including two steps (Text-to-UMLS and UMLS-to-ORDO). No annotation data are needed, through a UMLS extraction tool, SemEHR, and weak supervision (WS) based on customised rules and BERT-based contextual representations (see details on WS in Fig. 2). The admission ID and ICD-9 codes (linked with dotted lines) are only available for the MIMIC-III data. The lower, dotted lines show a baseline approach purely based on manual ICD codes, also enhanced with ontology matching. (Figure adapted from [7])

## Introduction

Text phenotyping is the task of extracting diseases or traits of patients from clinical notes, which can benefit a wide range of tasks like cohort selection, epidemiological research, and decision making for better clinical care. A particular set of human phenotypes are rare diseases: a rare disease is very uncommon, affecting 5 or fewer people in 10,000, but there are between 6,000 and 8,000 rare diseases and they collectively affect approximately 3.5-5.9% of the population (or 263-446 million persons) globally [1] (and over 1 in 17 people in the UK [2] and 8% of population in Scotland [3]) at some point in their lifetime. Compared to common diseases, rare diseases are usually not coded in a precise manner, this is partly because they are under-represented in the current, ICD-10 (International Classification of Diseases, version 10) terminologies [4, 5]. Detailed information about a patient is usually hidden in unstructured, clinical narratives. It is thus necessary to use clinical notes with Natural Language Processing (NLP) techniques to complement coded data to identify rare diseases in patients.

The main challenge for rare disease identification with NLP is the lack of annotated data for machine learning, especially deep learning. Deep learning models for clinical note classification tend to perform worse for infrequent diseases due to the lack of cases for training [6]. On the other hand, annotating a variety of rare diseases in clinical notes from scratch needs specific domain expertise. This also requires the manual annotation of a very large number of clinical notes to ensure enough cases for each rare disease, thus taking time and incurring considerable costs from a group of clinical experts.

We propose an ontology-driven and weakly supervised framework for rare disease identification from clinical notes, extending our previous work in [7] with further, detailed empirical analyses and external validation. Ontologies are essential for text phenotyping as they provide a curated list of terms of diseases and traits. Previous studies have used ontologies to estimate the frequency of rare diseases [8]. Our main ontology-driven framework is illustrated in Fig. 1.

We use Orphanet Rare Disease Ontology [9] as the list of vocabularies of rare diseases[1]. We then leverage the concepts and synonyms in Unified Medical Language System (UMLS) as an intermediary dictionary to extend matching terms and address the issue of name variation [12] in linking texts to rare diseases, e.g. "tracheobronchomalacia" for Williams-Campbell syndrome. The framework thus contains two integrated parts, *entity linking* (Text-to-UMLS) and *ontology matching* (UMLS-to-ORDO). Entity linking from mentions (or text fragments) to UMLS concepts is challenging due to the ambiguous mentions [8, 12], especially for abbreviations, e.g. "HD" which could mean Huntington Disease, Hemodialysis, or Hospital Day. String matching usually does not consider the complex contexts of a mention and can therefore result in many false positives. Machine learning

---

[1] We focus on the identification of diseases instead of the associated phenotypic abnormalities, therefore we chose ORDO instead of Human Phenotype Ontology (HPO) [10]. The overall ontology based and weak supervision framework can potentially be applied to HPO, given it being aligned to ORDO [11]. We leave the phenotypic abnormalities (in HPO) for future studies.

can be applied for the disambiguation of terms, but it needs abundant annotated training data, which are currently not available in the context of rare diseases.

We therefore propose a weakly supervised approach to filter out the false positives in entity linking. Weak supervision [13, 14] is a strategy to automatically create labelled training data using heuristics, knowledge bases, crowdsourcing, and other sources, to alleviate the burden and cost of annotation. We first use a string matching based named entity linking tool, SemEHR [15] (widely applied for text phenotyping in the UK [15–17], based on Bio-YODIE [18]) to generate candidate entity linking results, i.e. mentions and their UMLS concepts, from clinical notes; then, we propose to efficiently create weak training data of candidate mention-UMLS pairs of sufficient quality with two rules, mention character length, regarding ambiguous abbreviations, and "prevalence", regarding rare diseases. A phenotype confirmation model can thus be learned through contextual mention representations with domain-specific BERT models (e.g. Blue-BERT [19]) to capture the context under-lied in the texts to disambiguate the mention to improve entity linking. For UMLS-to-ORDO matching, we used the mappings in ORDO and corrected the wrong links by filtering ORDO concepts with a phenome type as an upper class in the ontology [9].

For our main experiments, we trained a weakly supervised phenotype confirmation model using the discharge summaries in the MIMIC-III dataset [20]. A large, weak entity linking dataset (of 127,150 candidate mention-UMLS pairs) was created for training. For evaluation, we annotated 1,073 mention-UMLS pairs as a gold-standard dataset. By filtering out the false positives, the proposed approach dramatically improved the precision and $F_1$ of the entity linking tool, SemEHR, with almost no loss of recall.

We further evaluated the phenotype confirmation models from discharge summaries to radiology reports in US MIMIC-III and UK NHS Tayside through either a direct transfer of the model or a weakly supervised re-training from new clinical notes. Almost perfect (100%) recall was achieved with a dramatic absolute increase of precision by over 30% to 50% with re-training and parameter tuning. This demonstrates that the approach can be efficiently adapted to identify rare disease phenotypes in another type of clinical notes and from another institution. Our annotated datasets on discharge summaries and radiology reports in MIMIC-III and our implementation of the overall approach are publicly available[2].

As far as we know, this is the first study on text phenotyping of rare diseases using weak supervision, with the application on clinical notes of different types and institutions. Our findings will shed light on using weakly supervised approaches and contextual representations for text phenotyping from clinical notes. The overall approach to identifying rare disease cohorts has the potential to support epidemiology and clinical decision making for better care.

## Background and related work

**Text phenotyping with ontologies.** Compared to the efficient and gradually economical genotyping (i.e. sequencing genomics information), *phenotyping* usually needs high-throughput computational approaches for the extraction of diseases and traits from electronic health records (EHRs) [21, 22]. Clinical codes (e.g. with International Classification of Diseases, ICD) are a common source typically used regarding their ease of retrieval for phenotyping. However, ICD codes are usually less specific to define nuanced diseases or traits (e.g. rare diseases [4]) and are likely to be incomplete or under-coded [23], which may cause erroneous and missing cases in phenotyping. An alternative source for phenotyping is free-text clinical notes in the EHRs. It is shown in a previous systematic review of cohort identification from EHRs [24] that text phenotyping (or case detection) achieves on average higher precision (or positive prediction value) and recall (or sensitivity) than code-based phenotyping, and combining both sources (texts and codes) achieved greatly improved phenotyping results. Text phenotyping also requires understanding the wider *contextual features* of the matched concepts, including negation (i.e. whether negated or hypothetical), experiencer (i.e. whether experienced by the patient or someone else), and temporality (i.e. whether historical) [16, 25]. These contextual features have been reasonably well detected with rule-based approaches, e.g. [25], and applied in Bio-YODIE and SemEHR, and more recently with neural network methods, e.g. in MedCAT [26].

Ontologies are essential for text phenotyping as they define the concepts and terms of diseases and traits. These concepts and terms are widely used to annotate clinical notes, i.e. match to text fragments or mentions [27] and to estimate rare diseases from texts [8]. The task to match ontology concepts (and their terms) to mentions is formally referred to as *entity linking.* One main issue of entity linking is entity ambiguity, where a mention could possibly denote different concepts or terms in an ontology [12]. Our work aims to improve entity linking with better disambiguation using *weak supervision* and *contextual mention representation*.

---

[2] https://github.com/acadTags/Rare-disease-identification

**Weak supervision.** Weak supervision [13, 14] is a strategy to efficiently create a large set of noisy labelled training data in a programmatical way using various sources containing heuristics and knowledge bases. The success of applying weak supervision in clinical NLP studies depends on two aspects, *data programming* and *data representation*, as suggested in [13]. Efficient *data programming* ensures that reliable weak data can be programmatically created for supervised learning. In clinical NLP, studies use lexical or concept filtering rules to create labelled data to extract nuanced categories (e.g. suicidal ideation [28] or lifestyle factors for Alzheimer's Disease [29]) from clinical texts. We extend over this line of research by using ontologies and a medical concept labelling tool with two specific rules to create reliable weak data to extract rare diseases. The second aspect is *data representation*, representing the contexts and semantics in the data into vectors in a high-dimensional space for subsequent steps in machine learning. For deep learning methods, previous studies [13, 29] proposed to use neural word embeddings and more recently using BERT [30] to represent the contexts of the textual data. We follow this direction to apply weak supervision with contextual representations for rare disease phenotyping.

**Contextual Representation.** The most significant, recent progress in NLP is the contextual representations pre-trained using Transformers [31] from a very large corpus [30]. The most representative contextual representation is BERT [30]. The pre-training task for BERT learns a masked language model with next sentence prediction, trained with a vast amount of curated texts on the Web (e.g. BookCorpus and English Wikipedia) using a 12 or 24 layered deep neural network mainly composed of multi-head self-attentions blocks. The learned parameters in the large neural network can then be applied to a wide range of downstream tasks, e.g. text classification, Named Entity Recognition, and question answering, with superior performance than the previous, task-specific models [30]. Contextual representations have been adapted to the clinical domain by pre-training using biomedical publications, clinical notes, and clinical ontologies. The notable models include but are not limited to BlueBERT [19] (BERT further pre-trained with PubMed abstracts and MIMIC-III clinical notes), PubMedBERT [32] (pre-trained from scratch with PubMed abstracts and full texts), SapBERT [33] (Pub-MedBERT further pre-trained with UMLS concepts), etc. We adapt the contextual representation methods for the mentions or text fragments to improve entity linking.

## Method

In this section, we will describe the ontology-driven method, the weak supervision for entity linking, contextual mention representation, and model training and inferencing.

### Entity linking and ontology matching

**Entity Linking.** Given a set of entities $E$ in an ontology and a collection of documents (e.g. clinical notes), entity linking aims to match a mention (or text fragment) $m$ to its corresponding entity $e \in E$ in the ontology [12]. The mention $m$ is a sequence of tokens in a document which potentially refers to one or more named entities and is usually identified in advance during the named entity recognition stage [12]. For Named Entity Recognition and Linking (NER+L) tools with a very large number of entities, e.g. Bio-YODIE [18], SemEHR [15], and Med-CAT [26], a mention $m$ is recognised at the same time when it is linked to a concept in an ontology; this is usually realised through string matching [18, 26].

We applied SemEHR, a medical NER+L tool widely deployed in Trusted Research Environments (or Data Safe Havens) and servers in the UK. Previously, high recall and $F_1$ (around 90%) were reported on sub-phenotyping with stroke from texts with SemEHR [17]. The output is a set of *mention-UMLS* pairs, where each mention is in a context window and with a name of the document structure (or the template section of the clinical note) if available. SemEHR adapts Bio-YODIE as its main NLP module, enhanced with a search interface and continuous learning functionalities based on users' feedback labels and rule-based and machine learning methods. Bio-YODIE can efficiently extract UMLSs from texts using a string matching based approach. When there is an ambiguous mention, time-efficient NER+L systems like Bio-YODIE mainly assume a corpus-based prior to assign the same, most frequent UMLS to the mention regardless of its context or surrounding texts [18]. This can result in many false positive phenotypes, mostly regarding the abbreviations in the clinical notes. For example in Table 1, none of the identified "HD" mentions indicate a type of disease, according to the context. While SemEHR has a continuous learning functionality to classify and correct the errors, the approach relies on users' feedback labels and requires time from clinical experts.

**Ontology Matching.** Another issue in entity linking is the variations of terms that may be missed in the process [12]. This can be addressed by using the rich term variations in the metathesaurus UMLS as an intermediary dictionary with *ontology matching* to match concepts in UMLS to ORDO. *Ontology matching* (or mapping) is the task of finding the correspondence
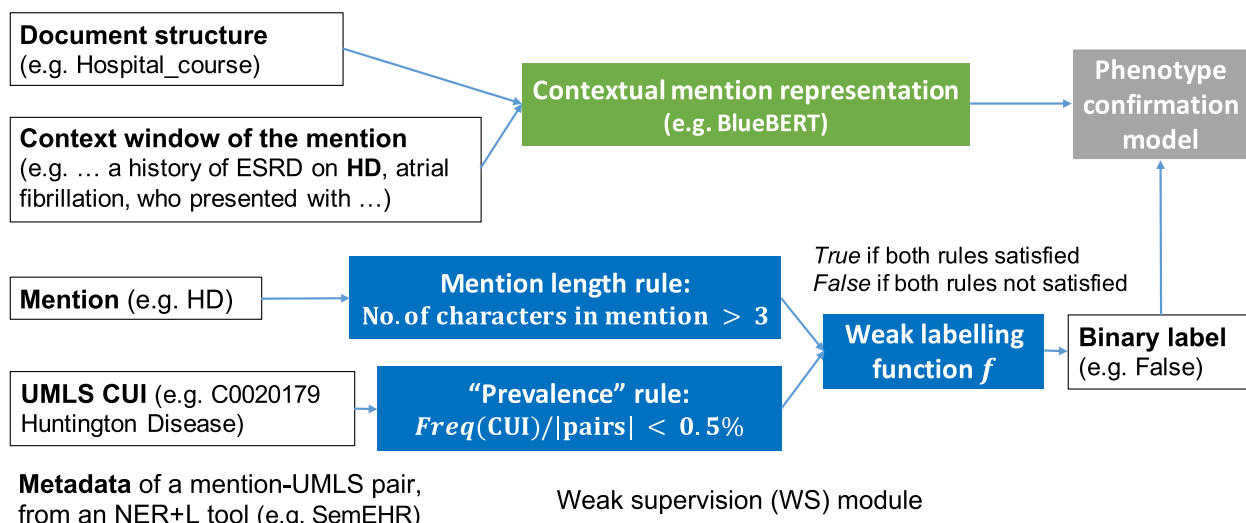
**Fig. 2** Weak supervision process for Text-to-UMLS linking. The left four white text boxes displayed the metadata (with examples) of a candidate mention-UMLS pair, identified by a Named Entity Recognition and Linking (NER+L) tool, SemEHR; the coloured text boxes in the middle show the contextual representation block (in `green`) and the rule-based weak data labelling (in `blue`. A binary label is then generated, which weakly estimates whether the candidate pair indicates a correct phenotype of the patient. A phenotype confirmation model (in `grey`) is then learned to select correct phenotypes from the pairs. (Figure adapted from [7]))

between two ontologies [34]. Each correspondence is represented as a triple $< e, e\prime, r >$, where $e$ and $e\prime$ denote an entity in the ontology $O$ and $O\prime$, respectively, and $r$ denotes a relation that holds between the two entities [35, p. 43]. The main form of an entity in an ontology is a *concept* or a *class*, denoted as $c \in C$ [35, p. 34]. In ORDO, the matching of an ORDO concept to UMLS and ICD-10 concepts are available as cross references [9], for example for Orphanet_3325 (Heparin-induced thrombocytopenia), there exist correspondences  $<$ Orphanet_3325, UMLS:C0272285, E $>$, where the relation E denote "Exact matching". We use E (Exact matching) or BTNT (ORDO's Broader Term maps to a Narrower Term) to ensure the matched term is a rare disease (and removed NTBT relations). We further added a rule ("isNotGroupOfDisorders") to filter out the Group of Disorders, e.g. Orphanet_181422 (Rare hyperlipidemia), which were mostly matched to a common disease in the UMLS, e.g. to C0020473 (hyperlipidemia). More details and examples of ontology matching are presented in Table S2-2 in Supplementary material 2.

### Weak supervision for phenotype confirmation model

To address the issue of ambiguous mentions, we propose weak supervision based on rules for labelled data creation with context mention embeddings for representation. When both data and representations are created, a classifier can be learned to decide whether a mention linked to UMLS in the context indicates a correct phenotype of the patient.

**Weakly Supervised Data Creation.** The idea in the weak data creation is to create rules that can complement the existing tool (e.g. SemEHR) to create reliable mention-UMLS pairs for training. The whole data creation process for weak supervision is described in the Algorithm 1. The candidate mention-UMLS pairs from an NER+L tool are denoted as a list of 5-element tuples $L$ (i.e. links), where each tuple includes a mention start position $m_{start}$, a mention end position $m_{end}$, a rare disease UMLS concept $c_{UMLS}^{rare}$, the context window of the mention $t$, and the name $s$ of the document structure where the mention is located. We propose two rules as functions on mention-UMLS pairs, *mention character length* rule, $\lambda_1$, and *"prevalence"* rule, $\lambda_2$, as shown in the blue blocks in Fig. 2. Given that abbreviations (like "HD" in Table 1) are usually ambiguous and falsely linked by the NER+L tools, the mention character length rule $\lambda_1$ satisfies when the mention has more than $l$ (default as 3) characters, i.e. $m_{end} - m_{start} > l$, otherwise as *False*. Given that rare diseases usually have a very low

**Table 1** Examples of false positives mention-UMLS pairs in entity linking identified from SemEHR and Bio-YODIE

| Mention in a context window | Meaning | *False positive* UMLS |
|---|---|---|
| temporary **HD** line was pulled. | Medical device | Huntington Disease |
| ... male with ESRD on **HD** ... | Haemodialysis | (C0020179) or |
| ... Asacol **HD** 800 mg Tablet ... | Medication | Hodgkin Disease |
| CT scan on **HD9** showed ... | Hospital Day | (C0019829) |

Each mention is bolded in its context window

prevalence [3, 36] and rare disease mentions usually have a low frequency in a consecutive sample of clinical notes, the "prevalence" rule $\lambda_2$ satisfies when the UMLS concept represents a very small percentage $p$ (default as 0.5%) in the whole number of candidate links $|L|$, i.e. $\frac{\text{Freq}(c)}{|L|} < p$, otherwise as *False*. This is an attempt to integrate an estimated epidemiological rule into weak supervision for text phenotyping.

---

**Require:** $T$, document set; $c_{UMLS} \in O_{UMLS}$, UMLS
        concepts and ontology; $c_{ORDO} \in O_{ORDO}$, ORDO
        concepts and ontology.
**Ensure:** $D_{weak}$, weakly labelled data
1  Initialise $D_{weak} \leftarrow \emptyset$;
2  $O_{\mathsf{UMLS}}^{\mathsf{rare}} = \{c_{UMLS} | c_{UMLS} \in < c_{ORDO}, c_{UMLS}, r >\}$;
3  $L \leftarrow \mathsf{SemEHR}(T, O_{\mathsf{UMLS}}^{\mathsf{rare}})$;
4  **for** *each* $< m_{start}, m_{end}, c_{UMLS}^{rare}, t, s > in\ L$ **do**
5  $\quad$ $\lambda_1 = m_{end} - m_{start} > l$;
6  $\quad$ $\lambda_2 = \frac{\mathsf{Freq}(c_{\mathsf{UMLS}}^{\mathsf{rare}})}{|L|} < p$;
7  $\quad$ **if** $\lambda_1$ *XNOR* $\lambda_2$ **then**
8  $\quad\quad$ $y_{weak} \leftarrow \lambda_1\ \mathsf{AND}\ \lambda_2$;
9  $\quad\quad$ $L_{weak} \leftarrow \mathsf{append}(L, y_{weak})$;
10 $\quad\quad$ $D_{weak} \leftarrow D_{weak} \cup L_{weak}$;
11 $\quad$ **end**
12 **end**

---

**Algorithm 1** Weakly supervised data creationThe final rule-based weak labelling function $\lambda$ is defined as *True* (i.e, mention-UMLS indicates a correct phenotype of the patient) when both rules $\lambda_1$ and $\lambda_2$ are satisfied, and as *False* when both rules are not satisfied. The data selection is equivalent to an XNOR logic operator (selected if and only if both rules are True or both are False) and the data labelling is equivalent to an AND operator of the rules. This ensures that only data that are consistently checked by both rules are weakly labelled. The binary weak label, $y_{weak} \in \{0, 1\}$, is then appended to each mention-UMLS pair to create the weakly labelled data $D_{weak}$.

The mention length threshold $l$ and the "prevalence" threshold $p$ are selected to ensure a sufficient amount of reliable, weak data generated. We empirically determine the best values of $l$ (as 3 or 4) and $p$ (as 0.005 or 0.01) based on the validation set or a small number of annotated data solely for evaluation (results on MIMIC-III discharge summaries in Table S1-1 in the Supplementary material 1).

**Contextual Mention Representation.** We use a clinically pre-trained BERT model (e.g. BlueBERT, as described in the related work) to represent the mention in its context window $t$ in the weakly labelled data $D_{weak}$. A BERT model can be succinctly described as the Eq. 1. We excluded layer normalisation, dropout, and other functions and parameters in the equations for simplicity.

The output $H^n \in R^{|\text{tokens}|, d}$ is a matrix that can be used as the layer for the subsequent task, where |tokens| is the length of sequence after tokenisation and $d$ denotes the dimensionality (usually 768 for BERTnorm and 1024 for BERTlarge). FFNN() is a feed-forward neural network of two linear transformations with a ReLU activation function in between, and MultiHead() is a multi-head self-attention layer that models multiple forms of alignment from the tokens to themselves; and the three inputs represent matrices of queries ($Q$), keys ($K$), and values ($V$), respectively, linearly transformed from $H^i$. We refer readers for the details of the Transformers and BERT architectures to [30, 31].

$$H^{i+1} = \text{FFNN}(\text{MultiHead}(W_Q H^i, W_K H^i, W_V H^i))$$
$$H^0 = \text{Embedding}(\text{Tokenize}(t))$$

$$(1)$$

The contextual understanding mainly comes from self-attention (as $\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$, where $d_k$ is a scaling factor) that captures the importance of every other token to each token. These parameters have been pre-trained based on massive corpora from general and medical domains. The hidden layers in BERT, $H$ can be used as *static* embeddings to represent a sequence. We extract the second-last layer $H^{n-1}$ in BERT as static embedding (or features) for the subsequent task, according to the results that $H^{n-1}$ has the best feature-based results among any single layers in $H$ for an NER task [30]. A plausible explanation for this is that the last layer is more biased towards the training loss (e.g. masked language model and next sentence prediction), while the second-to-last layer better represents the contextual information of the sentence.

The selection of the specific BERT model generally favours models pre-trained with in-domain (i.e. clinical) corpora [37] and is empirically based on results (e.g. $F_1$ scores) on the validation set. We will compare and analyse different BERT models in the experiments (see Table 4).

The overall weak supervision data representation and model training process is described in Algorithm 2. We use $H^{n-1} \leftarrow \text{BERT}(t)$ to denote the whole process above. Mean pooling, as empirically suggested in [38], is applied to create a final vector $v$. We define a contextual mention representation where only the tokens within the mention are included, i.e. $v \leftarrow \text{mean}(H^{n-1}[m_{start}^{\text{token}}, m_{end}^{\text{token}}])$. The start and end tokens' position of the mention $m_{start}^{\text{token}}$ and $m_{end}^{\text{token}}$ are derived based on the WordPiece tokenizer of the BERT model and the original position of the mention.

We also experimented with two encoding strategies, mention masking and using document structure name $s$ (see line 3 in Algorithm 2), that allow a more flexible

representation of the contexts. Non-masked encoding with document structures provided better results on the validation set (see Table S1-2 in Supplementary material 1).

**Model Training and Inference.** Finally, a *phenotype confirmation model* can be trained from the weakly labelled data. The contextual mention representation *v*, as static embedding, is fed into a binary classification model. We use logistic regression as the training model (in Train_and_validate() in Algorithm 2), which is similar to adding a feed-forward layer on top of the static pre-trained layer in BERT with sigmoid activation. We also compared this static embedding approach to fine-tuning the whole BERT model in the experiments.

---

**Require:** $< m_{start}, m_{end}, c_{\text{UMLS}}^{\text{rare}}, t, s, y_{weak} > \in D_{weak}$
**Ensure:** $M_{weak}$, the phenotype confirmation model
1  Initialise $X_{weak} \leftarrow \emptyset, Y_{weak} \leftarrow \emptyset$;
2  **for** $each < m_{start}, m_{end}, c_{UMLS}^{rare}, t, s, y_{weak} > in\ D_{weak}$ **do**
3  $\quad| \quad t \leftarrow \text{concatenate}(t, [\text{SEP}], s)$, if $s$ is not *null*;
4  $\quad| \quad H^{n-1} \leftarrow \text{BERT}(t)$;
5  $\quad| \quad m_{start}^{\text{token}}, m_{end}^{\text{token}} \leftarrow \text{tokenize}(t, m_{start}, m_{end})$;
6  $\quad| \quad v \leftarrow \text{mean}(H^{n-1}[m_{start}^{\text{token}}, m_{end}^{\text{token}}])$;
7  $\quad| \quad X_{weak} \leftarrow X_{weak} \cup v$;
8  $\quad| \quad Y_{weak} \leftarrow Y_{weak} \cup y_{weak}$;
9  **end**
10  $M_{weak} \leftarrow \text{Train\_and\_Validate}(X_{weak}, Y_{weak})$;

---

**Algorithm 2** Weakly supervised data representation and model trainingThe inference stage is succinctly defined in Eq. 2. We use SemEHR to extract candidate mention-UMLS pairs from a clinical note *d*. We then transform each instance into a contextual mention representation (see line 3-6 in Algorithm 2), denoted as the function $V_{\text{BERT}}()$. After selecting the patients' phenotype in $O_{\text{UMLS}}^{\text{rare}}$ with $M_{weak}$, we can then use the correspondence between UMLS and ORDO, denoted as $OM_{U \rightarrow O}$, to obtain the final set of rare disease phenotypes $C_{ORDO}^{d}$ as concepts in ORDO.

$$C_{ORDO}^{d} = OM_{U \rightarrow O}(M_{weak}(V_{\text{BERT}}(\text{SemEHR}(d, O_{UMLS}^{rare})))) \quad (2)$$

## Experiments

We evaluated the above ontology-driven and weakly supervised algorithms on MIMIC-III discharge summaries and further validated the approach with MIMIC-III radiology reports and NHS Tayside brain imaging reports. For validation and testing, we manually annotated a small number of mention-to-UMLS pairs from each of the datasets. We present results on each part of the system, Text-to-UMLS and UMLS-to-ORDO. For Text-to-UMLS, we carried out extensive experiments to study the best combination of parameters in weak

labelling rules, the encoding strategies, with a comparison between weak and strong supervision. We then show the whole pipeline can support rare disease phenotyping by enriching the traditional method using ICD codes. Finally, we show that the proposed approach can easily generalise or be adapted to a new type of clinical note, radiology reports, in the same or another institution.

### Data processing and annotation

We evaluated the proposed NLP pipeline with three datasets in two healthcare institutions in the US and the UK. The main dataset we used was the discharge summaries (n=59,652) in MIMIC-III ("Medical Information Mart for Intensive Care") dataset [20], which contains clinical data from adult patients admitted to the ICU in the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. We were granted access to MIMIC-III through PhysioNet after completing the ethical training by the Collaborative Institutional Training Initiative program. MIMIC-III data are supposed to contain rich rare disease mentions, as a large number of rare diseases (especially genetic disorders) can lead to an ICU (intensive care unit) admission [36].

The manual ICD-9 codes (i.e. ICD-9-CM) of the MIMIC-III admissions allow us to compare code-based phenotyping with text phenotyping for rare diseases. We linked ICD-9 codes to ICD-10 codes using the matching from the Ministry of Health, New Zealand [39] and linked ICD-9 to UMLS codes based on the ICD-9 ontology in BioPortal [40], as shown in Fig. 1. We used ORDO version 3.0 (released 07/03/2020), which contained 14,501 concepts or classes related to rare diseases. We selected the ORDO concepts which have linkage to UMLS and ICD-10 in this study as this supports the interoperability (e.g. linking and traversing) among the clinical terminologies; this resulted in a set of 4,064 rare disease concepts[3]. We focus on this essential set of overlapped rare diseases and the coverage is improving as the mappings are being updated; we leave the ORDO concepts without both ICD-10 and UMLS linkage for future research.

After processing the discharge summaries with a SemEHR database instance[4] [15] with rule-based contextual filtering on negation and experiencer based on [25], we obtained 127,150 candidate mention-UMLS pairs

---

**Table 2** Statistics of Clinical Note Datasets with the Natural Language Processing Pipeline and Manual Annotations

|  | MIMIC-III Disch | MIMIC-III Rad | Tayside Brain Img |
|---|---|---|---|
| $|T|$ | 59,652 | 522,279 | 156,618 |
| $|D|$ | 127,150 | 109,096 | 7,761 |
| $|D_{weak+}|$ | 15,598 | 13,907 | 1,137 |
| $|D_{weak-}|$ | 74,217 | 65,171 | 2,898 |
| $|T_{RD}|$ | 37,110 | 73,589 | 7,321 |
| $|T_{RD}^{weak}|$ | 10,568 | 21,102 | 2,855 |
| $|T^{ann}|$ | 500 | 1,000 | 5,000 |
| $|D^{ann}|$ | 1,073 | 198 | 279+4 |
| $|T_{RD}^{ann}|$ | 312 | 145 | 273 |

$|T|$, number of documents; $|D|$, number of mention-UMLS pairs; $|D_{weak+}|$, $|D_{weak-}|$, number of weakly labelled positive and negative mention-UMLS pairs, respectively; $|T_{RD}|$, $|T_{RD}^{weak}|$, number of documents associated with one or more rare diseases detected by SemEHR and SemEHR+WS (i.e. further with weak supervision), respectively; $|T^{ann}|$, $|D^{ann}|$, $|T_{RD}^{ann}|$, number of documents sampled, number of mention-UMLS pairs sampled, and number of the sampled documents with one or more rare diseases identified by SemEHR, respectively. For Tayside data, 4 new positive mention-UMLS pairs in $|D_{ann}|$ were identified from the reports during the manual annotation

for the UMLS concepts linked to ORDO. After applying the weak labelling function with the two rules, we finally obtained 15,598 positive and 74,217 negative data, and 37,335 non-labelled data or mention-UMLS pairs.

We further applied the same preprocessing steps with the MIMIC-III radiology reports (n=522,279) and NHS Tayside brain imaging reports (n=156,618). MIMIC-III radiology reports are from the same institution and within the same time span as in MIMIC-III discharge summaries [20]. The Tayside data contain the routine brain MRI and CT scans from the National Health Service (NHS) Tayside Health Board, which have been applied in previous NLP research [17, 41]. We have received NHS Tayside Caldicott Guardian approval to use the anonymised brain imaging reports for this work.

The statistics of the three datasets, MIMIC-III discharge summaries ("Disch"), MIMIC-III radiology reports ("Rad"), and NHS Tayside brain imaging reports ("Tayside Brain Img"), with the Natural Language Processing pipeline and manual annotations, are presented in Table 2. MIMIC-III discharge summaries have *proportionally* more documents associated with at least one candidate rare diseases (identified by SemEHR), quantified by $\frac{|T_{RD}|}{|D|}$: 3.4 times more than MIMIC-III radiology reports and 13.3 times more than brain imaging reports in Tayside.

**Data Annotation.** For evaluation, we created a gold standard dataset of 1,073 candidate mention-UMLS-ORDO triplets (with each mention in a context window) generated by SemEHR and ontology matching in ORDO, from a set of 500 randomly sampled discharge summaries

from MIMIC-III, of which 312 (or 62.5%) discharge summaries have at least one *candidate* or potential "rare disease" mention. There were in total 95 types of rare disease associated with the mentions. Annotators were asked to label whether a mention-UMLS pair truly indicates a phenotype of the patient with an annotation guideline of detailed examples on hypothetical mentions. The mention-UMLS pairs were annotated by 3 domain experts, including two research fellows and one PhD student in Medical Informatics (MI). Based on the random 200 mention-UMLS pairs annotated by all 3 domain experts, the multi-rater Kappa value was 0.76. ORDO-to-UMLS concept matching was annotated by 2 domain experts (a research fellow and a PhD student in MI) and obtained a Kappa of 0.72. All contradictory and unsure annotations were resolved by a research fellow in biomedical science and MI. We used the first 400 data instances for model validation and the rest 673 for final testing.

To study how the model performs when it is directly transferred to or re-trained on other clinical notes, we further annotated 198 candidate mention-UMLS pairs in a sample of 1,000 radiology reports in MIMIC-III [20] and 279 candidate mention-UMLS pairs (with 4 new manually identified mentions) in a sample of 5,000 brain imaging reports in NHS Tayside [17]. Each dataset was annotated by two researchers in clinical science or MI with contradictions addressed by another researcher. The Kappa for MIMIC-III radiology reports and NHS Tayside reports were 0.88 and 0.86, respectively.

To note that the evaluation set is independent of the rules used for weak supervision, thus abbreviations and "popular" disease mentions were in the validation and testing data. This helps to test whether the phenotype confirmation model trained on the rule-based weakly labelled data can generalise to the full scenario that also contains the *unseen* mentions, which were filtered out during weak supervision.

### Implementation details

We used the open-source tool, bert-as-service[5] [42], built on Google AI's BERT implementation with Python Tensorflow[6] [30] for contextual mention representation. We tested a range of pre-trained BERT models (BERT, BlueBERT, PubMedBERT, and SapBERT) and selected BlueBERT-base [19] based on results on the validation set (see Table 4). We then trained a logistic regression model with the representations, with default configuration using scikit-learn [43] on the weakly labelled mention-UMLS pairs. We also implemented a word2vec embedding baseline with Gensim[7] and a BERT fine-tuning baseline with

---

**Table 3** Evaluation results of Text-to-UMLS linking on validation and testing data from MIMIC-III discharge summaries

| Text to UMLS | validation (n=142+/400) | | | test (n=187+/673) | | | test, *seen* in WS (n=80+/499) i.e. both rules [not] satisfied | | | test, *unseen* in WS (n=107+/174) i.e. only one rule satisfied | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| SemEHR [15] | 35.5 | **100.0** | 52.4 | 27.8 | **100.0** | 43.5 | 16.0 | **100.0** | 27.6 | 61.5 | **100.0** | 76.2 |
| + rules | 80.9 | 89.4 | 84.9 | 68.6 | 94.7 | 79.6 | **83.3** | 87.5 | **85.4** | 61.5 | **100.0** | 76.2 |
| + WS (rules+BERT) | **92.0** | 89.4 | **90.7** | 81.4 | 91.4 | **86.1** | **83.3** | 87.5 | **85.4** | 80.2 | 94.4 | 86.7 |
| + SS (anns+BERT) | - | - | - | 88.4 | 93.6 | 90.9 | 87.7 | 88.8 | 88.2 | 88.9 | 97.2 | **92.9** |

The column statistics (n=$N_+$+/*N*) show the number of positive data $N_+$ and all samples *N* in the dataset. SemEHR has a perfect reference recall, because all candidate mention-UMLS pairs were created using the tool. WS, weak supervision; SS, strong supervision. BlueBERT-base (PubMed+MIMIC-III) was used as the BERT model. The best scores, either or not considering strong supervision (SS), are bolded

**Table 4** Comparison among embeddings for weakly supervised Text-to-UMLS linking from MIMIC-III discharge summaries

| Text to UMLS | validation (n=142+/400) | | | test (n=187+/673) | | |
| --- | --- | --- | --- | --- | --- | --- |
| | P | R | $F_1$ | P | R | $F_1$ |
| Word2Vec-100 | 86.6 | 50.0 | 63.4 | 85.1 | 61.0 | 71.0 |
| Word2Vec-300 | 85.7 | 59.2 | 70.0 | 80.7 | 69.5 | 74.7 |
| Word2Vec-768 | 85.1 | 68.3 | 75.8 | 78.9 | 78.1 | 78.5 |
| BERT | 88.1 | 83.8 | 85.9 | 79.5 | 91.4 | 85.1 |
| PubMedBERT | 88.7 | 77.5 | 82.7 | 79.6 | 87.7 | 83.5 |
| SapBERT | 88.3 | 79.6 | 83.7 | **80.8** | 89.8 | 85.1 |
| BlueBERT-base | **90.1** | **89.4** | **89.8** | 80.4 | **92.0** | **85.8** |
| + fine-tuning | 84.6 | 88.7 | 86.6 | 73.5 | **92.0** | 81.7 |
| BlueBERT-large | 89.1 | 80.3 | 84.4 | 79.0 | 88.8 | 83.6 |

The column statistics (n=$N_+$+/*N*) show number of positive data $N_+$ and all samples *N* in the dataset. All word2vec-*k* embeddings were pre-trained from MIMIC-III discharge summaries, representing the mention as the averaged *k*-dimensional embedding of tokens in the context window. BERT models were used as static features (in the second-last layer) if not specified with "fine-tuning". The best scores, either or not considering strong supervision (SS), are bolded. We did not tune the optimal number of random weakly supervised training data for BlueBERT-base model (and all other models), thus its results were slightly below those in Table 3

Huggingface Transformers[8], with detailed parameters in *Embedding and Fine-tuning Settings* in Supplementary material 1. Our implementation of the experiments is available at https://github.com/acadTags/Rare-disease-identification.

As baselines, we compared the proposed approach ("SemEHR+WS") with SemEHR with the two rules only using an OR operation for the interest of higher recall ("SemEHR+rules"). We evaluated the baselines using precision, recall, and $F_1$ scores. Note that SemEHR had a reference recall of 100% as all candidate "rare disease" mentions were identified by SemEHR, which was the starting source for the annotations.[9]

We tuned the two parameters *l* and *p* (to 3 and 0.5%, respectively, if not specified) in the weak labelling rules (in Algorithm 1) by grid search based on the performance of validation data in MIMIC-III discharge summaries. The detailed parameter tuning results of *l* and *p* are in Table S1-1 in Supplementary material 1, *Weak Rule Parameter Tuning*. We also tuned the size of context windows (default as 5), which however, did not affect the performance, probably because our final representation was based on the position of the mention in the BERT layer (see line 6 in Algorithm 2). Also, we tuned the optimal number of random training mention-UMLS pairs needed (n=9k) based on the validation set, which had little impact on the results (<1% $F_1$ score).

In contrast to weak supervision (WS), we also provide results on strong supervision (SS), the traditional approach that trains a model from full manually labelled data. For MIMIC-III discharge summaries, we used the first 400 validation set in the full 1,073 mentions to train a model, $M_{strong}$, and test on the rest 673 mentions

---

[8] https://github.com/huggingface/transformers

[9] We also benchmarked the performance on MIMIC-III discharge summaries with recent NER+L tools, MedCAT [26] and Google Healthcare Natural Language API [44]. However, given that the results (especially recall) may favour SemEHR-based methods, we do not formally report the results of the two NER+L tools but make them available at https://github.com/acadTags/Rare-disease-identification/blob/main/supp-results.

Dong *et al. BMC Medical Informatics and Decision Making* (2023) 23:86

Page 10 of 17

with the same inferencing step in Eq. 2 but using $M_{strong}$ instead of $M_{weak}$. As manually labelled data are usually more reliable than weakly labelled data, the performance of strong supervision is considered as an upper bound in studies in weak supervision [45, 46].

We provide the results regarding each step in the pipeline (in Fig. 1), Text-to-UMLS linking and UMLS-to-ORDO matching, followed by the overall results on rare disease identification, Text-to-ORDO linking and admission-level ORDO concept prediction.

### Main results: text-to-UMLS linking

Table 3 shows the validation and testing results of Text-to-UMLS linking. With weak supervision (WS), the precision and $F_1$ of SemEHR has been greatly improved by around 55% and 40% absolute value, respectively, for both validation and testing data. Adding the two customised rules already improved the testing performance greatly by over 30% $F_1$ to SemEHR (as shown in SemEHR+rules), which validates the efficiency of the two proposed rules with the NER+L tool to create reliable weak annotations. Adding WS further outperformed the SemEHR+rules setting absolutely by around 10% precision (and 5% $F_1$), showing the usefulness of the contextual mention representation on filtering out false positives. The recall dropped slightly after introducing the two rules. This indicates the bias or noise in the rules with the current threshold ($p$ as 0.5% and $l$ as 3). Results with weak supervision are within a small gap of 5% $F_1$ of strong supervision with hand-labelled data. This, overall, demonstrates the potential of WS to improve text phenotype entity linking.

As a solid evaluation needs to assess the system with different biased test sets, we further split the testing data into those weakly labelled or unlabelled during the weak supervision. This helps analyse the impact of the rule-based weak supervision on the testing performance. "Seen" data mean that the mention-UMLS pairs were weakly labelled with $\lambda$, i.e. with both rules satisfied or both not satisfied (see line 7-11 in Algorithm 1); "unseen" data mean that only one of the rules was satisfied so that the data were not labelled in the process. WS improved the performance of SemEHR in both settings: while the weakly "seen" data were dramatically boosted by rules (by nearly 50% $F_1$), the "unseen" data were greatly improved (by 10% $F_1$) through the model generalised with contextual representations.

The "unseen" data can be further split into the case that only the mention character length rule ($\lambda_1$) or the prevalence rule ($\lambda_2$) is satisfied. The former, "unseen-$\lambda_1$" testing set (n=127, where 96 are positive mentions) has more mentions than the latter, "unseen-$\lambda_2$" (n=47, where 11 are positive mentions). SemEHR+WS obtained

substantially better P/R/$F_1$ performance on "unseen-$\lambda_1$" (84.1/99.0/90.9) than "unseen-$\lambda_2$" (46.2/54.5/50.0). This shows that mentions that are infrequent abbreviations (i.e., "unseen-$\lambda_2$") tend to be more challenging than frequent non-abbreviations (i.e., "unseen-$\lambda_1$"). In both scenarios, SemEHR+WS performed the best $F_1$ among the baselines except for strong supervision (SemEHR+SS). However, given that the number of testing samples is small, e.g. only 11 positive mentions for "unseen-$\lambda_2$", we do not formally report the breakdown of results to draw solid conclusions.

**Embedding and Encoding Strategies.** We compared the different *embedding* methods, including word embeddings and several BERT models pre-trained from different sources. Table 4 shows that contextual mention embeddings (e.g. with BERT, described in lines 4-6 in Algorithm 2) based methods greatly outperformed word embeddings, although increasing the dimensionality of word2vec embeddings improved their recall and $F_1$. For the contextual mention embeddings, we compared the vanilla BERT and representative pre-trained BERT models in the biomedical domain. We observed that Blue-BERT, pre-trained using the in-domain (or same-data), MIMIC-III clinical notes, outperformed the various BERT models only from general domains (e.g. BERT), biomedical publications (e.g. PubMedBERT), or clinical ontologies (e.g. SapBERT). This supports the use of in-domain pre-trained models, e.g. BlueBERT for the task, corroborating the conclusion from [37]. We also see that *neither* using fine-tuning (cf. feature-based) nor the large version of BlueBERT could improve the performance, which is probably because they introduce more learnable parameters (and a larger model size for BlueBERT-large), thus likely overfitting the weakly labelled data and underperforming on the real, testing data. We further compare the *encoding* strategies and found that non-masked encoding (with document structures) achieved the best $F_1$ scores on the validation data (see Table S1-2 in Supplementary material 1).

### UMLS-to-ORDO matching results

For UMLS-to-ORDO ontology matching, the original accuracy by the ORDO ontology was 87.4% (=83/95), if considering the repeated mentions in the whole 1073 evaluation data, the linking accuracy was 81.6% (=876/1073). The most frequent three false UMLS-to-ORDO mappings in ORDO were Hyperlipidemia (C0020473) to Rare hyperlipidemia (Orphanet_181422), Epilepsy (C0014544) to Rare epilepsy (Orphanet_101998), and Dyslipidemias (C0242339) to Rare dyslipidemia (Orphanet_101953), all linking a broader, common disease concept to its specific types in rare diseases under the phenome type or the upper class

Dong *et al. BMC Medical Informatics and Decision Making*        (2023) 23:86

Page 11 of 17

**Table 5** Results on rare disease identification (Text-to-ORDO) from MIMIC-III discharge summaries

| Text to ORDO | validation (n=64+/400) | | | test (n=82+/673) | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| SemEHR [15] | 18.7 | 95.3 | 31.3 | 13.9 | **92.7** | 24.1 |
| + rules | 53.9 | 75.0 | 62.7 | 49.0 | 86.6 | 62.6 |
| + WS (rules+BERT) | **67.6** | 75.0 | **71.1** | 64.7 | 80.5 | **71.7** |
| + SS (anns+BERT) | - | - | - | 73.3 | 80.5 | 76.7 |

The column statistics (n=$N_+$+/N) shows number of positive data $N_+$ and all samples $N$ in the dataset. WS, weak supervision; SS, strong supervision; anns, annotations. BlueBERT-base (PubMed+MIMIC-III) was used as the BERT model. The best scores, either or not considering strong supervision (SS), are bolded

[9] of *group of disorders* (Orphanet_557492). By filtering with ORDO's phenome type using "isNotGroupOfDisorders" (i.e. not under *group of disorders*), the UMLS-to-ORDO concept linking accuracy of the unique and repeated mentions was improved to 88.4% (from 87.4%) and 94.4% (from 81.6%), respectively, from the whole validation and testing data in the MIMIC-III discharge summaries.

### Overall mention-level and admission-level results

We finally obtained the mention-level results (Text-to-ORDO) based on the two parts of the system. The results, shown in Table 5, are consistent with Text-to-UMLS results. The overall metrics are lower than Text-to-UMLS results (71.7% vs 86.1% for testing $F_1$ score for WS) due to the imperfect matching between UMLS and ORDO. For a perfect UMLS-to-ORDO matching, the results of the Text-to-UMLS and Text-to-ORDO should be the same.

In the interest of detection of rare disease cases in admissions, we aggregated the mention-level results to admission-level results, where one admission may be associated with several unique rare diseases (each as a concept in ORDO). Thus, we report the standard micro-level label-based metrics for multi-label classification [47]. Micro-level metrics count each admission to a single ORDO concept as an instance and create a confusion matrix to calculate the precision, recall, and $F_1$ scores. We were also able to obtain ICD-based results purely based on ontology matching (from ICD-9 codes to ICD-10 or UMLS concepts then finally to ORDO concepts, as shown in Fig. 1). Admission-level results were generally consistent with mention-level (Text-to-UMLS and Text-to-ORDO) results. In terms of precision and $F_1$ score, weak supervision greatly improved the performance of SemEHR and outperformed other third-party tools, slightly below strong supervision, while the recall was the same for both WS and SS. We also obtained the admission-level results of ICD codes.

Admission-level results are presented in Table S1-3 in Supplementary Material 1. It is discovered that our NLP-based approach (SemEHR+WS) achieved better precision and $F_1$ scores than the code-based approach

(ICD). In terms of recall, ICD codes could only identify a few more rare diseases cases than SemEHR with weak supervision (e.g. 21 vs 20 out of 30 in the validation set and 36 vs 33 out of 42 in the test set, between ICD ∪ SemEHR+WS and SemEHR+WS). Note that this result may not be accurate as our annotation is based on the string matching based NER+L results from SemEHR, so the false positives from ICD-based cohorts may actually be true cases. Also, the number of positive data is much lower in admission-level results than in the mention-level (e.g. for testing data, 42 admissions vs. 187 mention-UMLS pairs). But nevertheless, our results show the essential role of free-texts and NLP methods for rare disease phenotyping; the results are consistent with the conclusion in [24] regarding general diseases.

### Error analysis

We breakdown the errors of the proposed approach ("SemEHR+WS") regarding Text-to-ORDO in MIMIC-III discharge summaries (see results in Table 5) in Fig. 3. There were altogether 91 errors (including 59 false positives and 32 false negatives), representing 8.5% from the 1,073 candidate mentions-UMLS-ORDO triplets, where 61 (or 5.7%) were from Text-to-UMLS stage and 30 (or 2.8%) *only* from the UMLS-to-ORDO stage (and 4 in both stages).

While rules are effective for WS, they may also introduce some bias. Over half 57.4% (or 35 of 61 errors) from the Text-to-UMLS side were likely due to the bias introduced from the weak rules, where the prediction was wrong when using the weak rules only. The other two main errors were either (i) semantic type errors (representing 26.2% or 16 out of 61), where the mention was a (negative) laboratory test (e.g. "legionella") or other unrelated types (e.g. "ENDO" as department name) instead of a disease, or (ii) diseases of hypothetical or negative contexts (represented 6.6% or 4 out of 61), which were not filtered out by the NER+L tool, SemEHR, and were also challenging for the annotators. The other errors (9.8%, 6 out of 61) were due to not enough information for human to decide or no exact reason found for the error. The issues above
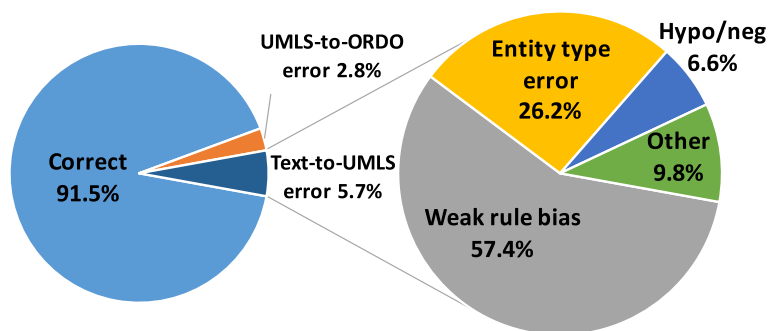
**Fig. 3** Error breakdown of Text-to-ORDO identification of 1,073 candidate mentions in MIMIC-III discharge summaries (Hypo/neg: Hypothetical or negation)
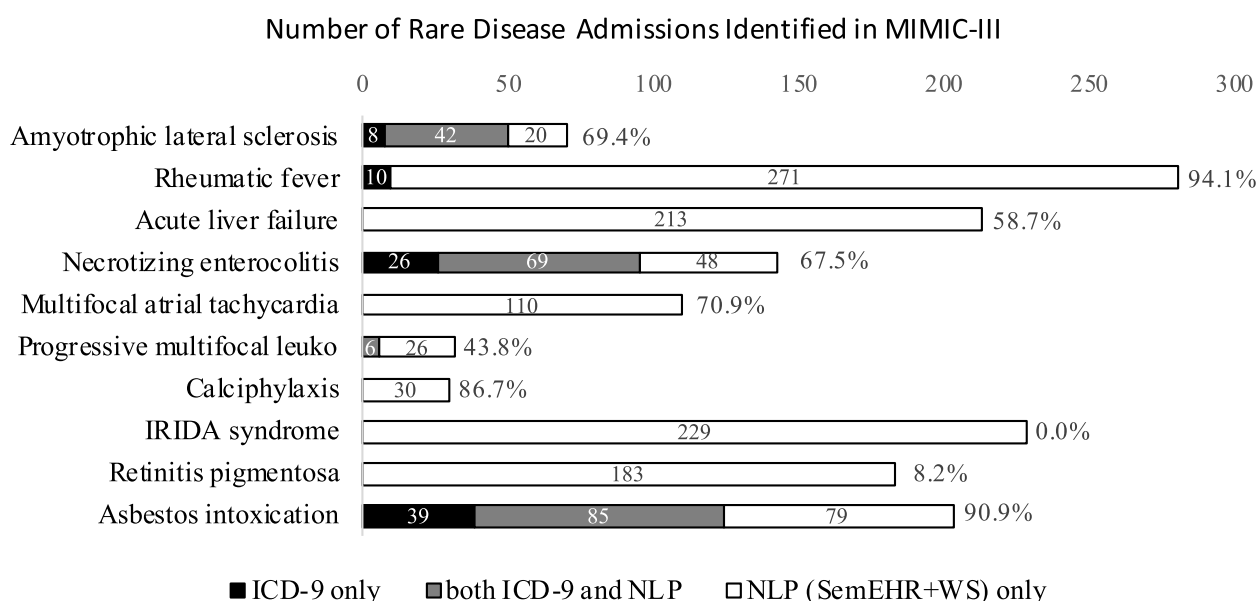


**Fig. 4** Number of rare disease patient stays from MIMIC-III (n=59,652): ICD (code-based) vs. NLP (text-based, with *weak* supervision), for 10 selected diseases. Admissions are split into those *only* identified through links from ICD-9 codes (in black), those *only* identified from free texts with weak supervision (NLP, in white), and the intersection of cases from both ICD-9 and NLP (in grey). The percentage after each horizontal bar shows the accuracy of NLP based on the manual assessment of the identified cases

may be addressed by combining WS with human-in-the-loop machine learning [48] with adaptive rules to improve the performance. The wrong UMLS-to-ORDO ontology mappings were due to the simple heuristic ("isNotGroupOfDisorders") which also filtered out correct mappings - this may be addressed when the official ontology matching is updated or by using a machine learning based system to correct the matching.

**NLP vs. ICD for rare disease phenotyping**
We applied the trained model and the whole pipeline to process all MIMIC-III discharge summaries (n=59,652) and compared the rare disease admissions identified from NLP and ICD. The NLP approach is the proposed ontology-driven and weakly supervised pipeline. For the ICD-based results, we combined the ICD-9 codes matched to either the UMLS or ICD-10 codes linked to ORDO (see Fig. 1).

Using our NLP-based pipeline, it is possible to greatly enrich the rare disease cases identified solely from ICD codes. For most (97.2%=453/466) types of the rare diseases, our approach mining free texts could enrich at least one (and usually many) potential rare disease case compared to the ICD-based approach. The results can be useful to identify potential cases for an alerting system for clinical care or a base for further refinement. Figure 4 shows the selected 10 rare diseases which were best predicted in the annotated 312

discharge summaries, however, since the support value was few (between 1 to 5) for each of the diseases in the admission-level evaluation, the results did not represent the predictions of the full 59k admission cases in MIMIC-III.

We thus further performed an extra manual evaluation to verify whether the rare disease cases identified by NLP were true phenotypes (or represented a current or past rare disease of the patient), as there was no gold reference standard. Five researchers (one in clinical science, one in biomedical science, and the remaining three in MI) screened the 1,428 cases or patient stays identified by NLP (WS or SS) regarding the 10 selected diseases, according to the definitions of the rare diseases in ORDO. The accuracy scores (the fraction of correct rare disease cases in all identified cases) of the weakly-supervised NLP-identified rare diseases are displayed after each horizontal bar in Fig. 4. We can see that NLP identified most rare diseases (6/10) with an accuracy score from around 70% to over 90%. For rheumatic fever, over 90% of the cases were true positives, except for a few hypothetical mentions or the subject being the patient's relative. Some examples are provided in Table S2-1 in Supplementary material 2. As rheumatic fever is usually a historical disease when the patient was a child, the disease was commonly not coded with ICD.

For certain rare diseases, the accuracy score from the manual evaluation was very low, e.g. 0.0% for IRIDA syndrome due to "microcytic anaemia" wrongly assigned as a synonym or an atom of C0085576 ("Iron-Refractory Iron Deficiency Anemia" or IRIDA) in the previous UMLS version (2019AA) in the Text-to-UMLS process, 8.2% and 43.8% for Retinitis Pigmentosa and Progressive Multifocal Leukoencephalopathy, respectively, due to the ambiguous meanings of their abbreviations ("RP" and "PML") and unseen in WS (with a low corpus-based prevalence below 0.5%). For Multifocal Atrial Tachycardia, the definition in ORDO is a neonatal disease, while its matched UMLS concept of the same name may also mean an adult disease. We also found difficulty in reaching a consensus in the annotation due to the vague definition of Acute Liver Failure in ORDO[10], for which we derived two distinct interpretations which were then reconciled by a senior clinician[11]. This analysis suggests that we should take the definitions into consideration in entity linking and ontology matching. We should also ensure that the

definitions used are appropriate for the clinical research question for people using the tools.

Although the accuracy scores were not perfect, for all diseases except IRIDA syndrome, NLP could still enrich the cases identified from ICD-9 after the manual check by the experts. We also find that with ICD codes, it is possible to find cases not identified by NLP as well, as shown in asbestos intoxication, necrotizing enterocolitis, etc., which may be related to the imperfect recall of the NLP model or the rare diseases being not (explicitly) mentioned in the clinical note. In general, the results above on *rare diseases* extend the conclusion of the previous survey in case detection [24] that NLP with free-texts can greatly enrich the information from ICD codes and the two sources complement each other. We further present the results of NLP with strong supervision in Fig. S1-1 in Supplementary material 1, which overall predicted fewer cases and resulted in better accuracy scores, but reflected the same picture as with weak supervision.

### Transfer and re-training with radiology reports

For external validation, we applied the proposed weak supervision pipeline and models to extract rare disease phenotypes from two datasets of radiology reports, US MIMIC-III radiology reports (n=520k) [20] and UK NHS Tayside brain imaging reports (n=156k) [17]. For each of the datasets, we selected a subset of clinical notes (1,000 for MIMIC-III and 5000 for Tayside), and obtained the candidate mention-UMLS pairs with SemEHR to be labelled for evaluation. The detailed data statistics are in Table 2. Based on the real-world practice of NLP, we consider two ways to apply the pipeline in Fig. 2: (i) model transfer and (ii) in-domain re-training. For model transfer, we directly applied our phenotype confirmation models, $M_{weak}$ (and $M_{strong}$), trained from MIMIC-III discharge summaries to the two new datasets; for in-domain re-training, we created weakly labelled training data from each new dataset and trained a data-specific phenotype confirmation model with Algorithms 1-2; we further tuned the parameters $p$ and $l$ in the weak labelling rules during re-training.

Table 6 shows the external validation results of the NLP pipeline with model transfer or in-domain re-training. We mainly present the Text-to-UMLS results, consistent with Text-to-ORDO results in Table S1-4 and admission-level results in Table S1-5 in Supplementary material 1. It is observed that directly applying a weak supervision model trained from another type of report (e.g. discharge summaries) could largely improve the precision and $F_1$ score of SemEHR, with a slight drop of recall from nearly 100% to over 90%. This transferability of models suggests that there are common linguistic patterns used in all types of clinical notes, even from different sources.

---

[10] https://www.ebi.ac.uk/ols/ontologies/ordo/terms?iri=http%3A%2F%2Fwww.orpha.net%2FORDO%2FOrphanet_90062

[11] Our two interpretations of acute liver failure differ most in the factors of drug use, alcohol abuse, virus infection, etc., that could contribute to the rarity of the disease but <u>not</u> specified in the definition from ORDO. We finally considered hepatitis virus or drugs as causes of acute liver failure as a rare disease, but removed cases of alcohol abuse.

**Table 6** External Validation Results on Radiology Reports from MIMIC-III and NHS Tayside

| Text to UMLS | MIMIC-III Radiology (n=53+/198) | | | Tayside Brain Imaging (n=79+/283) | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| SemEHR [15] | 26.7 | **100.0** | 42.2 | 26.9 | **94.9** | 41.9 |
| + WS (transfer) | 54.4 | 92.5 | 68.5 | 56.3 | 91.1 | 69.6 |
| + SS (transfer) | **89.4** | 79.2 | 84.0 | 69.0 | 62.0 | 65.3 |
| + rules (tuned) | 87.5 | 92.5 | 89.9 | 56.8 | 94.9 | 71.1 |
| + WS (in-domain) | 72.9 | 96.2 | 82.9 | 48.0 | 92.4 | 63.2 |
| + WS (+ tuning R) | 81.5 | **100.0** | 89.8 | 58.1 | **94.9** | 72.1 |
| + WS (+ tuning $F_1$) | 89.1 | 92.5 | **90.7** | **75.3** | 88.6 | **81.4** |

The column statistics (n=$N_+$+/$N$) show number of positive data $N_+$ and all samples $N$ in the dataset. WS, weak supervision; SS, strong supervision. The original parameters for WS were $p = 0.005$ and $l = 3$. The new parameters for best recall (R) were $p = 0.01$ and $l = 4$ and for best $F_1$ were $p = 0.0005$ and $l = 4$, for both datasets. For SemEHR+rules, we present the results of rules, where $p = 0.0005$ and $l = 4$, with an OR operation. The best scores for the metrics are bolded

The strong supervision model obtained a higher precision, but with a much lower recall (a drop of 20% to over 30% compared to SemEHR only) and thus may bear the risk of missing true positive mentions. Results from the in-domain re-training of models were much better than model transfer, as the former could bridge the linguistic gap between discharge summaries and radiology reports even for the same cohort or institution in MIMIC-III. We further tuned the weak labelling parameters to optimise the recall or $F_1$ score. A perfect or no loss of recall (100% or near 95%) was achieved on par with SemEHR and the precision was further improved compared to using the original parameters. Although the parameter tuning process was based on the full annotated data, this can be substituted by the inspection of a small number of data at the rule designing stage. Finally, we noticed that simply using rules (SemEHR+rules) with the best tuned parameters was highly effective, achieving better results than most evaluation settings, but still surpassed by the best tuned WS model, especially for the Tayside reports. The results between rules only and weak supervision were consistent with those of the discharge summaries in Table 3.

### Conclusion, discussion, and future studies
In this study, we proposed an ontology-driven and weakly supervised approach for rare disease phenotyping from clinical notes. Unlike the use of ontologies, weak supervision has not been well established in the clinical NLP domain. Our proposed weak supervised deep learning approach requires no human annotation and extends the paradigm from [13] on weak supervision for clinical texts, by introducing ontologies, named entity linking tools, and contextual representations. We designed two simple but effective rules (mention character length and corpus-based "prevalence") to create weakly labelled data regarding ambiguous abbreviations and rare entities. The

trained phenotype confirmation model effectively filtered out the false positives in the data with no (or a minimum) side effect on the true positives.

Traditional clinical NLP relies heavily on strong supervision with manually labelled data. However, with recent data-demanding methods like deep learning, it is time to consider to automatically create labelled data to train models, with the support of rules and resources like ontologies and NER+L tools. Our work on rare diseases provides empirical evidence for the task by applying a weakly supervised NLP pipeline on three clinical note datasets (one for discharge summaries and two for radiology reports) in two institutions in the US and the UK. The improvements on the precision were highly significant (by over 30% to 50% absolute score for Text-to-UMLS linking), with almost no loss of recall compared to the existing NER+L tool, SemEHR. Our study also demonstrates that NLP can complement traditional ICD-based approaches to better estimate rare diseases in clinical notes (see Fig. 4).

While our rule-based weak supervision does not require annotated data, it can bring bias or noise as no simple rule can perfectly predict the labels for a complex task. This bias, although not affecting most predictions for the testing data, was manifested in the slight drop of recall in Text-to-UMLS linking (Table 3). This loss of recall may be minimised through tuning the parameters in the weak labelling rule (e.g. relaxing the "prevalence" or mention length threshold, shown in Table 6), but needs a small set of annotated data or some manual inspection of the predictions. The mention character length rule may also be enhanced with accurate abbreviation expansion and disambiguation to retain abbreviations that are rare diseases. Besides, recent studies in the general NLP domain have begun tackling the bias of rules (with a rule-level attention mechanism [49]) or noise of weakly labelled data (with the estimation of data-level

confidence [50]). Also, we used a heuristic-based logic operation (as XNOR) to aggregate the two rules; future studies can explore more advanced aggregation methods (e.g., learning a label model [45, 46]).

As suggested in our results and other studies [45, 46], the current performance of the best weakly supervised methods is still below strong supervision. But the gap between the weak and strong supervision is small (within 5% $F_1$ score) and there is no difference in terms of recall. This shows that the expensive and time-consuming annotations for text phenotyping may be greatly reduced, substituted by an alerting system or manual screening based on the predictions of a weakly supervised NLP system. With a small number of annotated data for parameter tuning, both the precision and recall of our weak NLP model were further improved (see Table 6). This may suggest a future study to better use a small sample of annotated data with the weakly annotated data for semi-supervised learning to improve the performance.

There are still, however, some false positive mentions detected by the proposed NLP pipeline, as shown in our analyses of the prediction errors and the identified cohorts (in Figs. 3-4). Disambiguating entity types (especially for abbreviations) still remains a challenge for text phenotying. This suggests to potentially integrate word sense disambiguation to enhance the weak supervision approach, e.g., through more reliable weak data creation. Also, errors in identifying hypothetical and negation ("Hypo/neg") mentions suggest to separately model "Hypo/neg" in the classification, which can be learned with mentions beyond the scope of rare diseases. Furthermore, the complexities of linguistic patterns of a (rare) disease may still require better representations beyond the current context window and may need to be enhanced with ontology concepts. Our evaluation of the NLP-identified cases suggests modelling the semantics of the lexical definitions in ontologies (e.g. ORDO) to improve entity linking and ontology matching.

Also, we note that our work is highly dependent on existing ontologies and their available matchings to each other. We leveraged and validated the matching among ORDO, UMLS, ICD-10, and ICD-9. The current matchings are generally correct, but not perfect (e.g. 88.4% accuracy of matching between UMLS and ORDO). A more accurate matching among ontologies, potentially corrected with machine learning [51], will improve the performance of our pipeline. It is also possible to directly match texts to ORDO, which can include rare diseases not contained in UMLS, but this does not leverage the synonyms in UMLS that represent the name variation of rare disease entities. Also, our approach cannot identify emerging rare disease entities, not contained in the ontologies and not thus easily captured by SemEHR, which is the next, challenging direction for our study.

While we only enhanced SemEHR with the weakly supervised phenotype confirmation model, the approach can be adapted to improve other NER+L tools and models to support more accurate rare disease cohort selection and coding. Recently, more packages and environments (e.g. Snorkel [46], skweak [52]) have been created to apply weak supervision in general domain NLP practice. Thus, a promising future study is to adapt the current weak supervision infrastructures or the ideas behind them to the clinical NLP domain and establish best practices in the field; a recent work adapting Snorkel [46] is Trove [45], which has not yet been applied to the domain of rare diseases, that involves additional ontologies and their mappings.

Our work mainly focused on identifying rare disease concepts in the clinical notes, while other physical, behavioural, and physiological characteristics need to be identified so as to establish a clinical diagnosis of a rare disease. We also mainly focused on rare diseases as a whole and the approach can be applied to identify specific rare diseases. Future work needs to extract a wider set of information to enhance rare disease phenotyping, and to facilitate the development of risk prediction tools for rare diseases to support decision making during the COVID-19 pandemic and beyond [53, 54].

## Abbreviations

| | |
|---|---|
| BERT | Bidirectional Encoder Representations from TransformersZ |
| UMLS | Unified Medical Language System |
| ORDO | Orphanet Rare Disease Ontology |
| MIMIC-III | Medical Information Mart for Intensive Care |
| NHS | National Health Services |
| NLP | Natural Language Processing |
| NER+L | Named Entity Recognition and Linking |
| ICD | International Classification of Diseases |
| WS | Weak Supervision |
| SS | Strong Supervision |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-023-02181-9.

**Additional file 1.**

**Additional file 2.**

## Availability of data and materials
The MIMIC III datasets are available at https://mimic.physionet.org/ upon request after the ethical training. NHS Tayside data are not publicly available due to the privacy of patients and please refer to https://www.dundee.ac.uk/hic regarding further interest in the dataset. The rare disease mention annotations of MIMIC-III discharge summaries and radiology reports, along with the implementation of the approach, are available at https://github.com/acadTags/Rare-disease-identification.

## Declarations

### Ethics approval and consent to participate
We were granted access to MIMIC-III through PhysioNet after completing the ethical training in human research subject protections and HIPAA regulations, through the Collaborative Institutional Training Initiative program (https://physionet.org/about/citi-course/). We have also received NHS Tayside Caldicott Guardian approval (CSAppMW1758) to use the anonymised brain imaging reports for this work. All our methods were carried out in accordance with relevant guidelines and regulations. The approval of both MIMIC-III and NHS Tayside datasets allows us to carry out Natural Language Processing experiments on the reports. All reports have been de-identified and we do not identify any individual patients in the methods and experiments, thus the research is exempt from requiring informed consent from the patients according to the NHS Tayside Caldicott Guardian approval.

### Consent for publication
Not applicable.

### Competing interests
The authors declare that they have no competing interests.

### Author details
[1]Centre for Medical Informatics, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom. [2]Health Data Research UK, London, United Kingdom. [3]Department of Computer Science, University of Oxford, Oxford, United Kingdom. [4]Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom. [5]Institute of Health Informatics, University College London, London, United Kingdom. [6]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom. [7]Department of Computer Science, The University of Manchester, Manchester, United Kingdom. [8]Edinburgh Futures Institute, University of Edinburgh, Edinburgh, United Kingdom.

## References
1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165–73.
2. Department of Health & Social Care. The UK Rare Diseases Framework. 2021. https://www.gov.uk/government/publications/uk-rare-diseases-framework/the-uk-rare-diseases-framework. Accessed 8 May 2022.
3. Scottish Government. Illnesses and long-term conditions. 2021. https://www.gov.scot/policies/illnesses-and-long-term-conditions/rare-diseases/. Accessed 22 Mar 2021.
4. Richesson RL, Fung KW, Bodenreider O. Coverage of Rare Disease Names in Clinical Coding Systems and Ontologies and Implications for Electronic Health Records-Based Research. In:  Proceedings of the 5th International Conference on Biomedical Ontology. Houston: CEUR Workshop Proceedings (CEUR-WS.org); 2014. p. 78–80.
5. Bearryman E. Does your rare disease have a code? 2016. https://www.eurordis.org/news/does-your-rare-disease-have-code. Accessed 29 July 2021.
6. Dong H, Suárez-Paniagua V, Whiteley W, Wu H. Explainable Automated Coding of Clinical Notes using Hierarchical Label-wise Attention Networks and Label Embedding Initialisation. J Biomed Inform. 2021;103728. https://doi.org/10.1016/j.jbi.2021.103728.
7. Dong H, Suárez-Paniagua V, Zhang H, Wang M, Whitfield E, Wu H, Rare disease identification from clinical notes with ontologies and weak supervision. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Online: IEEE; 2021. p. 2294–8.
8. Kahn Jr CE. An Ontology-Based Approach to Estimate the Frequency of Rare Diseases in Narrative-Text Radiology Reports. Stud Health Technol Inf. 2017;245:896–900. MEDINFO 2017: Precision Healthcare through Informatics.
9. Vasant D, et al. ORDO: an ontology connecting rare disease, epidemiology and genetic data. In Bio-Ontology @ ISMB 2014. 2014. p. 1-4. https://www.researchgate.net/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data.
10. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am J Hum Genet. 2015;97(1):111–24. https://doi.org/10.1016/j.ajhg.2015.05.020.
11. Maiella S, Olry A, Hanauer M, Lanneau V, Lourghi H, Donadille B, et al. Harmonising phenomics information for a better interoperability in the rare disease field. European Journal of Medical Genetics. 2018;61(11):706–714. Focus on rare disease research projects supported by the E-Rare ERA-Net program. https://doi.org/10.1016/j.ejmg.2018.01.013.
12. Shen W, Wang J, Han J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Trans Knowl Data Eng. 2015;27(2):443–60. https://doi.org/10.1109/TKDE.2014.2327028.
13. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inf Decis Making. 2019;19(1):1. https://doi.org/10.1186/s12911-018-0723-6.
14. Ratner A, Varma P, Hancock B, Ré C, other members of Hazy Lab. Weak Supervision: A New Programming Paradigm for Machine Learning. 2019. http://ai.stanford.edu/blog/weak-supervision/. Accessed 13 Mar 2021.
15. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc. 2018;25(5):530–7. https://doi.org/10.1093/jamia/ocx160.
16. Wu H, Hodgson K, Dyson S, Morley K, Ibrahim Z, Iqbal E, et al. Efficiently Reusing Natural Language Processing Models for Phenotype Identification in Free-text Electronic Medical Records: Methodological Study. JMIR Med Inf. 2019;7(4):e14782:1-14.
17. Gorinski PJ, Wu H, Grover C, Tobin R, Talbot C, Whalley H, et al. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. arXiv preprint arXiv:1903.03985. 2019;Comment: 8 pages, presented at HealTAC 2019, Cardiff, 24-25/04/2019.
18. Gorrell G, Song X, Roberts A. Bio-yodie: A named entity linking system for biomedical text. arXiv preprint arXiv:1811.04860. 2018.

19. Peng Y, Yan S, Lu Z. 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. Florence: Association for Computational Linguistics; 2019. p. 58–65.

20. Johnson AEW, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):1–9. https://doi.org/10.1038/sdata.2016.35.

21. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inf Assoc. 2013;20(e2):e206–11. https://doi.org/10.1136/amiajnl-2013-002428.

22. Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc. 2013;20(e2):e253–9. https://doi.org/10.1136/amiajnl-2013-001945.

23. Searle T, Ibrahim Z, Dobson R. Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset. In: Proceedings of BioNLP. Online: Association for Computational Linguistics. 2020. p. 76–85. https://doi.org/10.18653/v1/2020.bionlp-1.8.

24. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23(5):1007–15. https://doi.org/10.1093/jamia/ocv180.

25. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biom Inform. 2009;42(5):839–851. Biomedical Natural Language Processing. https://doi.org/10.1016/j.jbi.2009.05.002.

26. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021;117:102083. https://doi.org/10.1016/j.artmed.2021.102083.

27. Kersloot MG, van Putten FJP, Abu-Hanna A, Cornet R, Arts DL. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semant. 2020;11(1):14. https://doi.org/10.1186/s13326-020-00231-z.

28. Cusick M, Adekkanattu P, Campion TR, Sholle ET, Myers A, Banerjee S, et al. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. J Psychiatr Res. 2021;136:95–102.

29. Shen Z, Schutte D, Yi Y, Bompelli A, Yu F, Wang Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. BMC Med Inf Decis Making. 2022;22(1):1–11.

30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT. Minneapolis, Minnesota: Association for Computational Linguistics. 2019. p. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach: NeurIPS Proceedings; 2017. p. 5998–6008.

32. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare. 2021;3(1). https://doi.org/10.1145/3458754.

33. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-Alignment Pretraining for Biomedical Entity Representations. In: Proceedings of NAACL-HLT. Online: Association for Computational Linguistics. 2021. p. 4228–4238. https://doi.org/10.18653/v1/2021.naacl-main.334.

34. Noy NF. Ontology Mapping. In: Staab S, Studer R, editors. Handbook on Ontologies. International Handbooks on Information Systems. Berlin, Heidelberg: Springer. 2009. p. 573–590. https://doi.org/10.1007/978-3-540-92673-3_26.

35. Euzenat J, Shvaiko P. The Matching Problem. In: Ontology Matching. Berlin, Heidelberg: Springer Berlin Heidelberg. 2013. p. 25–54. https://doi.org/10.1007/978-3-642-38721-0_2.

36. Textoris J, Leone M. Genetic Aspects of Uncommon Diseases. In: Leone M, Martin C, Vincent JL, editors. Uncommon Diseases in the ICU. Cham: Springer International Publishing; 2014. p. 3–11. https://doi.org/10.1007/978-3-319-04576-4_1.

37. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks.

In: Proceedings of ACL. Online: Association for Computational Linguistics. 2020. p. 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740.

38. Ma X, Wang Z, Ng P, Nallapati R, Xiang B. Universal text representation from bert: An empirical study. arXiv preprint arXiv:1910.07973. 2019.

39. Ministry of Health NZ. Mapping between ICD-10 and ICD-9. 2000. https://www.health.govt.nz/nz-health-statistics/data-references/mapping-tools/mapping-between-icd-10-and-icd-9. Accessed 30 Apr 2021.

40. NCBO BioPortal. International Classification of Diseases, Version 9 - Clinical Modification. 2021. https://bioportal.bioontology.org/ontologies/ICD9CM. Accessed 30 Apr 2021.

41. Sykes D, Grivas A, Grover C, Tobin R, Sudlow C, Whiteley W, et al. Comparison of rule-based and neural network models for negation detection in radiology reports. Nat Lang Eng. 2021;27(2):203–24. https://doi.org/10.1017/S1351324920000509.

42. Xiao H. Serving Google BERT in Production using Tensorflow and ZeroMQ. 2019. https://hanxiao.io/2019/01/02/Serving-Google-BERT-in-Production-using-Tensorflow-and-ZeroMQ/. Accessed 25 Apr 2021.

43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

44. Bodnari A. Healthcare gets more productive with new industry-specific AI tools. 2020. https://cloud.google.com/blog/topics/healthcare-life-sciences/now-in-preview-healthcare-natural-language-api-and-automl-entity-extraction-for-healthcare. Accessed 15 Mar 2021.

45. Fries JA, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun. 2021;12(1):1–11.

46. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid training data creation with weak supervision. VLDB J. 2020;29(2):709–30.

47. Gibaja E, Ventura S. A Tutorial on Multilabel Learning. ACM Comput Surv. 2015;47(3). https://doi.org/10.1145/2716262.

48. Monarch RM. Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Shelter Island, NY: Manning Publications Company; 2021. Version 11, MEAP Edition (Manning Early Access Program).

49. Karamanolakis G, Mukherjee S, Zheng G, Awadallah AH. Self-Training with Weak Supervision. In: Proceedings of NAACL-HLT. Online: Association for Computational Linguistics. 2021. p. 845–863. https://doi.org/10.18653/v1/2021.naacl-main.66.

50. Jiang H, Zhang D, Cao T, Yin B, Zhao T. Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data. In: Proceedings of ACL-IJCNLP. Online: Association for Computational Linguistics. 2021. p. 1775–1789. https://doi.org/10.18653/v1/2021.acl-long.140.

51. Kolyvakis P, Kalousis A, Smith B, Kiritsis D. Biomedical ontology alignment: an approach based on representation learning. J Biomed Semant. 2018;9(1):1–20.

52. Lison P, Barnes J, Hubin A. skweak: Weak Supervision Made Easy for NLP. In: Proceedings of ACL-IJCNLP: System Demonstrations. Online: Association for Computational Linguistics. 2021. p. 337–346. https://doi.org/10.18653/v1/2021.acl-demo.40.

53. Zhang H, Thygesen J, Wu H. Increased COVID-19 related mortality rate for patients with rare diseases: a retrospective cohort study with data from Genomics England. Lancet. 2021;398:S95. Public Health Science 2021. https://doi.org/10.1016/S0140-6736(21)02638-6.

54. Zhang H, Thygesen JH, Shi T, Gkoutos GV, Hemingway H, Guthrie B, et al. Increased COVID-19 mortality rate in rare disease patients: a retrospective cohort study in participants of the Genomics England 100,000 Genomes project. Orphanet J Rare Dis. 2022;17(1):1–7.

## Publisher's Note