Supplementary Material for “Ontology-Based and Weakly Supervised Rare Disease Phenotyping from Clinical Notes”

SA. Data Statistics

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 S1.

TABLE S1: Statistics of Clinical Note Datasets with the Natural Language Processing Pipeline and Manual Annotations

	MIMIC-III Disch	MIMIC-III Rad	Tayside Brain Img
$\lvert T\rvert$	59,652	522,279	156,618
$\lvert D\rvert$	127,150	109,096	7,761
$\lvert D_{weak+}\rvert$	15,598	13,907	1,137
$\lvert D_{weak-}\rvert$	74,217	65,171	2,898
$\lvert T_{RD}\rvert$	37,110	73,589	7,321
$\lvert T_{RD}^{weak}\rvert$	10,568	21,102	2,855
$\lvert T^{ann}\rvert$	500	1,000	5,000
$\lvert D^{ann}\rvert$	1,073	198	279+4
$\lvert T_{RD}^{ann}\rvert$	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, resp.; $|T_{RD}|$, $|T_{RD}^{weak}|$, number of documents associated to one or more rare diseases detected by SemEHR and SemEHR+WS (i.e. further with weak supervision), resp.; $|T^{ann}|$, $|D^{ann}|$, $|T_{RD}^{ann}|$, 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.

SB. Setting of NER+L tools for benchmarking

MedCAT We use the official version of MedCAT[1] [1] with their vocabulary and concept database (storing concepts and their embeddings). Similar to our approach using string matching (as in SemEHR [2]) and with a weakly supervised model for entity disambiguation (using exact matching to the canonical name of an entity as the rule), MedCAT can match to nested mentions and learn concept embeddings based on the context window for disambiguation. The concept embeddings are updated incrementally each time based on the embeddings of the sampled positive and negative *contexts*. Context embeddings are modelled as an average of word embeddings of tokens in a context window. The word embedding used in MedCAT is Word2vec [3], empirically outperforming the static clinically pre-trained BERT embedding in the official experiments [1]. There are three types of models for MedCAT, small, medium, and large. Our best results on the validation set were achieved by either the small or the medium model, with the confidence score threshold as 0.2, and not using the contextual features or the “meta-annotations” [1], e.g. negation.

Google Healthcare Natural Language API (GHNL API) GHNL API[2] [4] identifies clinical entities from texts and links them to UMLS and other ontologies. The contextual filtering settings were “certainly assessment” no less than “SOMEWHAT LIKELY” and “subject” as “PATIENT”.

For both tools, we assume that the mention-UMLS pair is predicted as *True* if the same UMLS concept is detected as the one in the annotated data. We found nearly no affect ($< 0.05\%$ F_1) applying a tolerance value (as 5) of mention positions when we matched mention spans detected by the tools to those in the annotated data, thus we reported result of exact matching, i.e., no tolerance. To note that both tools are being maintained and updated and that we conducted the experiments with GHNL API in March 2021 and with MedCAT in June 2021.

SC. Weak Rule Parameter Tuning

The results of parameter tuning for weak labelling rules regarding F_1, recall, and precision scores, are displayed in Table S2. We tuned the possible values of $p \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ and $l \in \{2, 3, 4\}$ through a grid search, and selected the model based on recall and F_1 scores in Text-to-UMLS linking. For MIMIC-III discharge summaries, the results were based on the 400 validation set of manually annotated mention-UMLS pairs. The parameter p controls the corpus-based “prevalence” of the disease concept, which is related to epidemiological information, e.g., the actual prevalence of a rare disease in the cohort. A higher p resulted in more disease concepts selected, thus higher recall but generally less precision. The parameter l controls the mention length, a key threshold to filter out abbreviations, which are usually ambiguous in their meanings. A higher l thus generally resulted in a higher precision but lower recall. We observed that a corpus-based “prevalance” threshold of 0.005 and a mention character length threshold of 3 resulted in the best F_1 score for the dataset. We thus recommend to set $p \in \{0.005, 0.01\}$ and $l \in \{3, 4\}$ and used p as 0.005 and l as 3 for MIMIC-III discharge summaries.

[1] `https://github.com/CogStack/MedCAT`

[2] `https://cloud.google.com/healthcare/docs/concepts/nlp`

TABLE S2: F_1, Precision (P), and Recall (R) scores with respect to the weak rule parameters p and l in Text-to-UMLS linking for MIMIC-III discharge summaries (with the highest F_1 score in bold)

	l = 2			l = 3			l = 4		
	F_1	P	R	F_1	P	R	F_1	P	R
p = 0.0001	73.7%	64.1%	86.6%	59.3%	86.5%	45.1%	64.3%	87.8%	50.7%
p = 0.0005	73.7%	59.6%	96.5%	82.6%	91.5%	75.4%	79.8%	91.0%	71.1%
p = 0.001	72.8%	57.3%	100.0%	80.8%	91.2%	72.5%	79.8%	91.0%	71.1%
p = 0.005	71.5%	55.7%	100.0%	**89.8%**	90.1%	89.4%	87.5%	91.5%	83.8%
p = 0.01	71.0%	55.0%	100.0%	89.7%	87.3%	92.3%	88.5%	90.4%	86.6%
p = 0.05	63.7%	46.7%	100.0%	64.0%	47.0%	100.0%	64.3%	47.3%	100.0%
p = 0.1	60.0%	42.9%	100.0%	60.0%	42.9%	100.0%	60.0%	42.9%	100.0%

SD. Results on Different Encoding Strategies

The first encoding strategy is *mention masking*, whether or not to mask the mention in the full context window. The intuition behind this is to explore the potential of a language model to confirm a phenotype solely based on the surrounding context but not the mention itself.

The second encoding strategy is *using document structure names* (or template section names) to enhance local context. If the document structure name s is available in the dataset, we add s before the context window t with a separation token [SEP] in between.

Results on the different encoding strategies for Text-to-UMLS linking in MIMIC-III discharge summaries are displayed in Table S3. Non-masked encoding achieved significantly better results than masked encoding. Using document structure names further boosted recall scores on the validation and the test set. We used non-masked encoding (with document structure names for MIMIC-III discharge summaries only) for data representation.

TABLE S3: Comparison among encoding strategies for weakly supervised Text-to-UMLS linking on MIMIC-III discharge summaries

	validation set (n=142+/400)			test set (n=187+/673)		
Text to UMLS	P	R	F_1	P	R	F_1
non-M	89.9	87.3	88.6	**81.3**	90.9	**85.9**
non-M+DS	**90.1**	**89.4**	**89.8**	80.4	**92.0**	85.8
M	86.5	63.4	73.2	78.6	61.0	68.7
M+DS	86.4	62.7	72.7	78.0	62.6	69.4

M denotes mention masking and non-M denotes no mention masking applied. DS denotes using document structure names. The non-M+DS model was trained on the full set of weakly labelled data, without tuning the optimal number of data, thus slightly below results in Table II in the paper. BlueBERT-base (PubMed+MIMIC-III) was used to encode the text sequences.

SE. NLP with Strong Supervision vs. ICD for Admission-level Rare Disease Phenotyping

Figure S1 shows the results of the NLP pipeline with strong supervision compared to ICD codes for admission-level rare disease phenotyping. The results were generally consistent with the weak supervision approach (in Figure 4 in the paper) that NLP-based results greatly complement the code-based rare disease cohort. Generally, a higher accuracy with a less number of admissions was predicted by strong supervision compared to weak supervision (e.g. the accuracy was 25.5% or 14/55 predicted by "Retinitis Pigmentosa" for strong supervision, compared to 8.2% or 15/183 predicted by weak supervision).

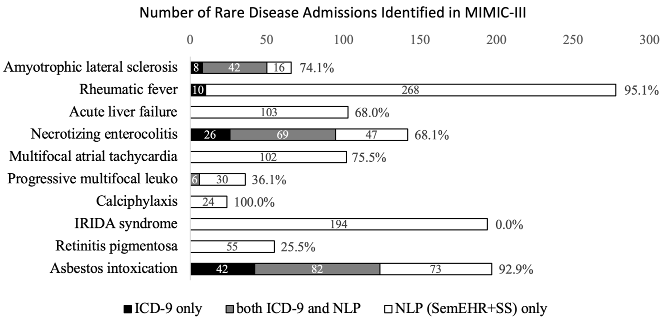

Fig. S1: Number of rare disease patient stays from MIMIC-III (n=59,652): ICD (code-based) vs. NLP (text-based, with *strong* supervision). The 10 rare diseases are the same as those presented for weak supervision in Figure 4 in the paper. Admissions are split into those *only* identified through links 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.

SF. Overall Admission-level and Mention-level Results

Table S4 shows the admission-level rare disease phenotyping results for MIMIC-III discharge summaries, analysed in Section IV.E in the paper.

Table S5 and S6 show the overall mention-level (Text-to-ORDO) and admission-level results of two radiology report datasets in the US (MIMIC-III) and the UK (NHS Tayside). For Tayside data, the recall was lower as we manually identified new rare disease mentions that were not included in the candidate mentions from SemEHR. Weak supervision (WS) achieved significantly better recall than transferring the SS model in results from both Tables. The code-based approach (ICD) also did not show an advantage in identifying more rare disease admissions (see recall, R), and overall performance (see F_1), comparing ICD or "ICD ∪ SemEHR+WS" with the (best) SemEHR+WS setting in Table S6 and Table S4, but the results may be biased towards methods adapting SemEHR as it was used as a starting source to create candidate mentions for the manual annotation.

TABLE S4: Micro-level results of admission-level rare disease identification for MIMIC-III discharge summaries

	validation			test		
	(n=30+/117∗55)			(n=42+/192∗82)		
Admission to ORDO	P	R	F_1	P	R	F_1
GHNL API [4]	45.7	70.0	55.3	40.4	54.8	46.5
MedCAT [1]	47.9	76.7	59.0	37.7	69.0	48.7
SemEHR [2]	15.4	**93.3**	26.4	12.7	**95.2**	22.3
+ rules	39.2	66.7	49.4	38.9	88.1	54.0
+ WS (rules+BERT)	**57.1**	66.7	**61.5**	**49.3**	78.6	**60.6**
+ SS (anns+BERT)	-	-	-	**61.1**	78.6	**68.7**
ICD	56.2	30.0	39.1	27.3	21.4	24.0
ICD ∪ SemEHR+WS	50.0	70.0	58.3	40.4	85.7	55.0
ICD ∪ SemEHR+SS	-	-	-	45.9	81.0	58.6

The micro-level metric counts each admission and an associated ORDO concept (or an admission-ORDO pair) as a single instance. The column statistics (n=N_++/$N_d * N_l$) show the number of positive data N_+, admissions (or discharge summaries) N_d, and unique candidate rare diseases (or ORDO concepts) N_l in the dataset. WS, weak supervision; SS, strong supervision; anns, annotations. BlueBERT-base (PubMed+MIMIC-III) was used as the BERT model. ICD denotes the approach to match ICD-9 codes to ORDO concepts. The union sign (∪) denotes merging and de-duplicating the cases identified from the two methods. Precision (P) and F_1 for ICD-based methods may be lower than actual values, as all candidate mentions were from SemEHR.

TABLE S5: Results on rare disease identification (Text-to-ORDO) for MIMIC-III and Tayside radiology reports

	MIMIC-III Radiology			Tayside Brain Imaging		
	(n=46+/198)			(n=42+/283)		
Text to ORDO	P	R	F_1	P	R	F_1
SemEHR	22.9	**93.5**	36.8	13.1	**78.6**	22.4
+ WS (transfer)	48.8	84.8	61.9	31.4	76.2	44.4
+ SS (transfer)	86.5	69.6	77.1	**53.2**	59.5	56.2
+ rules (tuned)	84.8	84.8	84.8	31.4	76.2	44.4
+ WS (in-domain)	68.3	89.1	77.4	26.4	**78.6**	39.5
+ WS (+ tuning R)	78.2	**93.5**	85.1	32.4	**78.6**	45.8
+ WS (+ tuning F_1)	**86.7**	84.8	**85.7**	46.3	73.8	**56.9**

The column statistics (n=N_++/N) shows the number of positive data N_+ and the overall number of 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, rules were aggregated with an OR operation and $p = 0.0005$ and $l = 4$.

SG. Examples of Rare Disease Text Phenotyping

Table S7 shows some selected prediction errors and a few correct predictions. The first four examples are the false positives selected in the evaluation data for the weak supervision model due to semantic type errors, hypothetical contexts, or other issues, analysed in Section IV.F in the paper. The last five examples are those selected from the identified rare disease cohort for Retinitis Pigmentosa and Rheumatic Fever, analysed in Section IV.G in the paper. Synonyms in UMLS could help identify some entity variations, e.g. “tracheobronchomalacia” for Williams-Campbell syndrome, and “acute rheumatic fever” for Rheumatic fever, but also introduces false positives especially regarding abbreviations, e.g. “EMA” and “RP”. The complex context in the clinical notes, including the relative’s diseases or hypothetical mentions, although only representing a small part of cases, were still challenging for the NLP pipeline (SemEHR+WS), as these were not explicitly considered in our weakly supervised training process. We also note that there were errors in parsing the document structure name through regular expressions in SemEHR, which might affected the predictions.

TABLE S6: Micro-level results of admission-level rare disease identification for MIMIC-III and Tayside Radiology Reports

	MIMIC-III Radiology			Tayside Brain Imaging		
	(n=29+/145∗43)			(n=41+/273∗65)		
Admission to ORDOs	P	R	F_1	P	R	F_1
SemEHR	19.4	**93.1**	32.1	12.8	**78.0**	22.0
+ WS (transfer)	38.7	82.8	52.7	30.7	75.6	43.7
+ SS (transfer)	**83.3**	69.0	75.5	**53.2**	61.0	56.8
+ rules (tuned)	80.0	82.8	81.4	30.7	75.6	43.7
+ WS (in-domain)	59.5	86.2	70.4	25.8	**78.0**	38.8
+ WS (+ tuning R)	71.1	**93.1**	80.6	31.7	**78.0**	45.1
+ WS (+ tuning F_1)	82.8	82.8	**82.8**	46.3	75.6	**57.4**
ICD	46.4	44.8	45.6	-	-	-
ICD ∪ SemEHR+WS	51.9	**93.1**	66.7	-	-	-

The micro-level metric counts each admission and an associated ORDO concept (or an admission-ORDO pair) as a single instance. The column statistics (n=N_++/$N_d * N_l$) show the number of positive data N_+, the number of admissions (or discharge summaries) N_d, and the number of candidate rare diseases (or ORDO concepts) N_l 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, rules were aggregated with an OR operation and $p = 0.0005$ and $l = 4$. The union sign (∪) denotes merging and de-duplicating the cases identified from the two methods. For ICD ∪ SemEHR+WS, the WS model was “in-domain + tuning R”, the one re-trained with in-domain data and optimised recall. Precision (P) and F_1 for ICD-based methods may be lower than actual values, as all candidate mentions were from SemEHR.

SH. Ontology Matching from ORDO to ICD-9

Table S8 shows 10 examples of rare disease concepts and their ontology matching from ORDO to UMLS, ICD-10, and finally to ICD-9. The rare diseases are the same as those presented in Figure 4 in the paper and Figure S1.

References

[1] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, and R. J. Dobson, “Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit,” *Artificial Intelligence in Medicine*, vol. 117, p. 102083, 2021.

[2] H. Wu, G. Toti, K. I. Morley, Z. M. Ibrahim, A. Folarin, R. Jackson, I. Kartoglu, A. Agrawal, C. Stringer, D. Gale, G. Gorrell, A. Roberts, M. Broadbent, R. Stewart, and R. J. Dobson, “SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research,” *Journal of the American Medical Informatics Association*, vol. 25, no. 5, pp. 530–537, 01 2018.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[4] A. Bodnari, *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 Mar. 15, 2021).

[5] 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 Apr. 30, 2021).

[6] NCBO BioPortal, *International Classification of Diseases, Version 9 - Clinical Modification*, 2021, https://bioportal.bioontology.org/ontologies/ICD9CM (accessed Apr. 30, 2021).

TABLE S7: Examples of wrong and correct rare diseases identified by SemEHR with the weak supervised phenotype confirmation model from MIMIC-III discharge summaries

ROW_ID	Document Structure	Text (with **mention** in bold)	UMLS	ORDO	Pred	Label	Potential Reason
26825	pertinent_results (should be pathology)	Pathology: ...Immunostains for cytokeratin AE1/3 and CAM 5.2, CD-68, CD-79a, CD-138, S-100, LCA absorbed CEA, **EMA**, CD34, CD31, TTF-1, actin, desmin, MNF-116, calcitonin, and thyroglobulin are negative...	C0268596	26791 Multiple acyl-CoA dehydrogenase deficiency	T	F	negation with a long context, ambiguous mention (EMA as epithelial membrane antigen), and semantic type error (negative test)
869	Hospital_course	Brief Hospital Course: ## Dyspnea - ...DFA for flu was negative; urinary **legionella** antigen was also negative...	C0023241	549 Legionellosis	T	F	semantic type error with negation (negative test)
8960	History_of_Past_Illness	Past Medical History: 1. Diagnosed in his early years with bilateral uveitis, clinically had bilateral uveitis significant with loss of vision and **sarcoid** floaters in both eyes…	C0036202	797 Sarcoidosis	T	F	not enough information (sarcoid floater not necessary means sarcoid)
46361	pertinent_results (should be impression)	IMPRESSION: ...Of note prior chest CT scans have findings suggesting a propensity to **tracheobronchomalacia**, as well as moderately severe emphysema....	C0340231	411501 Williams-Campbell syndrome	T	F	hypothetical context
48161	Admission_Medications	Medications on Admission: Hydrochlorothiazide/Triamterene 37.5mg/25mg, Protonix 40mg daily, Advair 250mcg/50 mcg daily, Singulari 10mg daily, Flovent, Flonase, Nasonex, [**Doctor First Name **] 60mg daily, Migquin 65mg-325mg-100mg daily, Gabapentin 100mg daily, Nortriptyline 20mg qhs, Lexapro 15mg daily, Concerta 18mg dialy, Clonazepam 2mg qhs, Restasis, Systane and Lotemax ophthalmologic drops, Vitamin A palmitate 100,000 units 1.5 tablets dialy for **retinitis pigmentosa**, acetaminophen, tums, Mylanta, OTC Prilosec prn	C0035334	791 Retinitis Pigmentosa	T	T	correct
26351	Hospital_course	...Her ASA continued to be held due to the **RP** bleed but was restarted after 48 hrs of stable Hct…	C0035334	791 Retinitis Pigmentosa	T	F	ambiguous abbreviation (Retroperitoneal bleeding)
12659	History_of_Past_Illness	Past Medical History: PMHx: ... 7. h/o of **rheumatic fever** with Sydenham's chorea…	C0035436	3099 Rheumatic fever	T	T	correct
20984	Hospital_course	The patient never reported any pharyngitis, but given his complaints of diffuse arthralgias, myalgias, migrating neuropathic pain, there was some concern of **rheumatic fever**, as the patient had 2 ASO screens performed which were both negative.	C0035436	3099 Rheumatic fever	T	F	hypothetical context
11568	basic (should be family history)	FAMILY HISTORY: ...2) His mother has an enlarged heart which may be secondary to a history of **acute rheumatic fever**...	C0035436	3099 Rheumatic fever	T	F	a relative's disease

Prediction errors are coloured with red in "Pred" (third-last) column. For columns "Pred" and "Label", "T" means that the prediction or gold is *True* and "F" indicates *False*. The wrongly parsed document structure names in the second column are marked with corrected ones in the form of "(should be XXX)".

TABLE S8: Ontology concept matching among ORDO, UMLS, ICD-10, and ICD-9 based on publicly available sources

ORDO	ORDO Preferred Label	UMLS	ICD-10	ICD-9-NZ (from ICD-10)	Preferred Label	ICD-9-BP (from UMLS)	Preferred Label
803	Amyotrophic lateral sclerosis	C0002736	<G122	-	-	335.20	Amyotrophic lateral sclerosis
3099	Rheumatic fever	C0035436	>I011, >I00, >I010, >I012, >I018, >I019	3911, 390, 3910, 3912, 3918, 3919	Acute rheumatic endocarditis; Rheumatic fever without mention of heart involvement; Acute rheumatic pericarditis; Acute rheumatic myocarditis; Other acute rheumatic heart disease; Acute rheumatic heart disease, unspecified	390-392.99	ACUTE RHEUMATIC FEVER
90062	Acute liver failure	C0162557	<K720	-	-	-	-
391673	Necrotizing enterocolitis	C0520459	=P77	7775	Necrotizing enterocolitis in newborn	-	-
3282	Multifocal atrial tachycardia	C0221158	<I471	-	-	-	-
217260	Progressive multifocal leukoencephalopathy	C0023524	=A812	0463	Progressive multifocal leukoencephalopathy	46.3	Progressive multifocal leukoencephalopathy
280062	Calciphylaxis	C0006666	<E835	-	-	-	-
791	Retinitis pigmentosa	C0035334	<H355	-	-	-	-
209981	IRIDA syndrome	C0085576	<D508	-	-	-	-
2302	Asbestos intoxication	C0003949	<J61	-	-	501	Asbestosis

=, >, and < in ORDO-to-ICD-10 mappings (all from ORDO) indicate exact, broader-to-narrower, and narrower-to-broader matching, respectively. Narrower-to-broader matching (<) from ORDO to ICD-10 was not used for phenotyping, as it may result in common or non-rare diseases' ICD codes. All ORDO-to-UMLS mappings (all from ORDO) indicate exact matching (=). "ICD-9-NZ" denotes the set of ICD-9 codes linked from ICD-10 codes using the matching from the Ministry of Health, New Zealand [5]. "ICD-9-BP" refers to the set of ICD-9 codes linked from UMLS based on the ICD-9-CM ontology (version 2020AB) in BioPortal [6].