# Extended supplementary material - Benchmarking results with recent Named Entity Recognition and Linking (NER+L) tools on rare disease identification

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]

Full list of author information is available at the end of the article

This is an extended supplementary material for our work in [1] on rare disease identification. We benchmarked the performance on MIMIC-III discharge summaries which two recent, representative NER+L tools for text-to-UMLS linking, (i) MedCAT[1] [2], using string matching enhanced with disambiguation based on neural word embeddings, which outperformed several established NER+L tools and has been applied to hospitals in the UK and (ii) Google Healthcare Natural Language API (GHNL API)[2], an enterprise-oriented, third-party tool designed for clinical texts, released on Nov 2020 [3]. We tuned the best parameters for both tools with the validation set to optimise $F_1$ score. Note that both tools are being updated and our experiments were carried out in March 2021 for MedCAT and June 2021 for GHNL API. We refer readers to [1] for details on the main methods and the data. The data annotations, predictions, and code are available at `https://github.com/acadTags/Rare-disease-identification`.

Given that the focus of this study is on weak supervision rather than using a NER+L tool, we did not include other established tools in the benchmarking, e.g. cTAKES and MetaMap, considering that they are technically similar to SemEHR and Bio-YODIE (mainly based on string matching) and were previously compared with MedCAT and Bio-YODIE in [4, 2].

*Setting of NER+L tools for benchmarking*

**MedCAT** We use the official version of MedCAT[3] [2] with their vocabulary and concept database (storing concepts and their embeddings). Similar to our approach using string matching (as in SemEHR [5]) 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 [6], empirically outperforming the static clinically pre-trained BERT embedding in the official experiments [2]. 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" [2], e.g. negation.

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

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

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

**Google Healthcare Natural Language API (GHNL API)** GHNL API[4] [3] 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 effect ($< 0.05\%$ $F_1$) 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 the results of exact matching, i.e., no tolerance.

**SemEHR-enhanced methods** As baselines for ablation studies, we compared (i) the proposed approach ("SemEHR+WS", see main paper in [1]) with (ii) SemEHR with the two rules only using an OR operation for the interest of higher recall ("SemEHR+rules"). We evaluated the approach 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; recall (R) may favour SemEHR-based methods, but precision (P) is fairly comparable across systems.

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 with the same inferencing step 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 [7, 8].

We provide the results regarding each step in the pipeline, 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 E-1 Extended evaluation results of Text-to-UMLS linking on validation and testing data from MIMIC-III discharge summaries**

| | 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 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Text to UMLS | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| GHNL API [3] | 78.9 | 81.7 | 80.3 | 75.3 | 78.1 | 76.6 | 54.3 | 62.5 | 58.1 | **94.1** | 89.7 | 91.9 |
| MedCAT [2] | 83.3 | 91.5 | 87.2 | 70.3 | 87.2 | 77.8 | 56.6 | 75.0 | 64.5 | 81.7 | 96.3 | 88.4 |
| SemEHR [5] | 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.

Table E-1 shows the extended 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

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

**Table E-2 Extended results on rare disease identification (Text-to-ORDO) from MIMIC-III discharge summaries**

| | validation (n=64+/400) | | | test (n=82+/673) | | |
|---|---|---|---|---|---|---|
| Text to ORDO | P | R | $F_1$ | P | R | $F_1$ |
| GHNL API [3] | 47.6 | 60.9 | 53.4 | 44.2 | 51.2 | 47.5 |
| MedCAT [2] | 59.6 | 82.8 | 69.3 | 48.8 | 76.8 | 59.7 |
| SemEHR [5] | 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.

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). Also with WS, the overall approach outperforms the recent NER+L tools, GHNL API and MedCAT, by about 6-11% absolute *precision* (if not considering the recall which favours SemEHR-based systems regarding data annotation). Even though GHNL API and MedCAT were not specifically designed for rare diseases, our results show the importance of weak supervision to enhance an NER+L tool with customised rules and contextual mention representation to outperform both, most recent, off-the-shelf tools. 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.

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; "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. We also see that MedCAT and GHNL API achieved slightly better results in linking mentions to UMLS for the weakly "unseen" testing set (but worse for the "seen" testing set).

*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 extended results, shown in Table E-2, are consistent with Text-to-UMLS results. Table E-3 further shows the extended admission-level rare disease phenotyping results for MIMIC-III discharge summaries.

**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. Dong, H., Suárez-Paniagua, V., Zhang, H., Wang, M., Casey, A., Davidson, E., Chen, J., Alex, B., Whiteley, W., Wu, H.: Ontology-based and weakly supervised rare disease phenotyping from clinical notes. arXiv preprint arXiv:2205.05656 (2023)
2. Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A.A., Roberts, A., Bendayan, R., Richardson, M.P., Stewart, R., Shah, A.D., Wong, W.K., Ibrahim, Z., Teo, J.T., Dobson, R.J.B.:

**Table E-3 Extended micro-level results of admission-level rare disease identification for MIMIC-III discharge summaries**

| | validation (n=30+/117∗55) | | | test (n=42+/192∗82) | | |
|---|---|---|---|---|---|---|
| Admission to ORDO | P | R | $F_1$ | P | R | $F_1$ |
| GHNL API [3] | 45.7 | 70.0 | 55.3 | 40.4 | 54.8 | 46.5 |
| MedCAT [2] | 47.9 | 76.7 | 59.0 | 37.7 | 69.0 | 48.7 |
| SemEHR [5] | 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.

Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. Artificial Intelligence in Medicine **117**, 102083 (2021). doi:10.1016/j.artmed.2021.102083

3. 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 Mar. 15, 2021)
4. Gorrell, G., Song, X., Roberts, A.: Bio-yodie: A named entity linking system for biomedical text. arXiv preprint arXiv:1811.04860 (2018)
5. Wu, H., Toti, G., Morley, K.I., Ibrahim, Z.M., Folarin, A., Jackson, R., Kartoglu, I., Agrawal, A., Stringer, C., Gale, D., Gorrell, G., Roberts, A., Broadbent, M., Stewart, R., Dobson, R.J.: 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 **25**(5), 530–537 (2018). doi:10.1093/jamia/ocx160
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
7. Fries, J.A., Steinberg, E., Khattar, S., Fleming, S.L., Posada, J., Callahan, A., Shah, N.H.: Ontology-driven weak supervision for clinical entity classification in electronic health records. Nature communications **12**(1), 1–11 (2021)
8. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. The VLDB Journal **29**(2), 709–730 (2020)