Efficient Diagnosis Assignment Using Unstructured Clinical Notes

Louis Blankemeier
Stanford University
lblankem@stanford.edu

Jason Fries
Stanford University
jfries@stanford.edu

Robert Tinn
Microsoft Health AI
robert.tinn@microsoft.edu

Sam Preston
Microsoft Health AI
sam.preston@microsoft.edu

Nigam Shah
Stanford University
nigam@stanford.edu

Akshay Chaudhari
Stanford University
akshaysc@stanford.edu

Abstract

Electronic phenotyping entails using electronic health records (EHRs) to identify patients with specific health outcomes and determine when those outcomes occurred. Unstructured clinical notes, which contain a vast amount of information, are a valuable resource for electronic phenotyping. However, traditional methods, such as rule-based labeling functions or neural networks, require significant manual effort to tune and may not generalize well to multiple indications. To address these challenges, we propose *HyDE* (hybrid diagnosis extractor). HyDE is a simple framework for electronic phenotyping that integrates labeling functions and a disease-agnostic neural network to assign diagnoses to patients. By training HyDE's model to correct predictions made by labeling functions, we are able to disambiguate hypertension true positives and false positives with a supervised area under the precision-recall curve (AUPRC) of 0.85. We extend this hypertension-trained model to zero-shot evaluation of four other diseases, generating AUPRC values ranging from 0.82 - 0.95 and outperforming a labeling function baseline by 44 points in F1 score and a Word2Vec baseline by 24 points in F1 score on average. Furthermore, we demonstrate a speedup of $> 4\times$ by pruning the length of inputs into our language model to $\sim 2.3\%$ of the full clinical notes, with negligible impact to the AUPRC. HyDE has the potential to improve the efficiency and efficacy of interpreting large-scale unstructured clinical notes for accurate EHR phenotyping.

1 Introduction

The widespread adoption of electronic health records (EHRs) by health systems has created vast clinical datastores. One of the essential steps in utilizing these data is identifying patients with specific clinical outcomes and the timing of these outcomes, through a process called electronic phenotyping (Banda et al., 2018). Electronic phenotyping is critical for using EHR data to support clinical care (Kaelber et al., 2012; LePendu et al., 2012), inform public health decision-making (Dubberke et al., 2012), and train predictive models (Chaves et al., 2021; Blankemeier et al., 2022; Steinberg et al., 2021, 2023; Lee et al., 2022).

Electronic phenotyping is a complex task that involves combining structured data (e.g. lab results and codes) with unstructured data (e.g. clinical notes). Rule-based heuristics can be applied to structured data. However, the unstructured nature of information rich (Kern et al., 2006; Wei et al., 2012; Martin-Sanchez and Verspoor, 2014) clinical notes makes phenotyping based on these notes particularly challenging.

Several solutions exist for electronic phenotyping using unstructured clinical notes (Peng et al., 2018; Fries et al., 2021; Zhang et al., 2021a,b), but lack convenience for generalizing to new conditions. For example, labeling functions that consist of rules authored by domain experts are interpretable and readily shared without compromising data privacy, but can be laborious to create. Neural networks (NNs) that are trained to identify specific diseases can eliminate the need for handcrafted labeling functions and often provide more accurate results. However, NNs require extensive manual labeling time and often generalize poorly to diseases not seen during training.

To address this, we introduce HyDE (hybrid diagnosis extractor). HyDE is a simple approach to electronic phenotyping that combines the strengths of labeling functions and neural networks and allows for generalization to new diseases with minimal overhead.

Our key contributions are as follows:

1. We demonstrate that our model effectively discriminates between true cases of hypertension and false positives generated by labeling functions, as demonstrated by a supervised area under the precision recall curve (AUPRC) of

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 485–494

0.85. This same model achieves AUPRCs of 0.90, 0.82, 0.84, and 0.95 in zero-shot evaluations for *diabetes, osteoporosis, chronic kidney disease*, and *ischemic heart disease*, respectively. HyDE outperforms a labeling function baseline by 44 points in F1 score and a Word2Vec baseline (Mikolov et al., 2013b,a) by 24 points in F1 score on average across seen and unseen diseases.

2. HyDE requires minimal setup. The labeling functions used in HyDE can be simple, reducing the manual effort often required to design labeling functions with high precision and recall.

3. HyDE is computationally efficient, as only small portions of a subset of clinical notes need to be passed through the neural network for processing, thus minimizing the computational resources required to run HyDE on large datasets. We show that pruning the length of the inputs by $4\times$ to just 2.3% of the full clinical notes impacts performance by an average of only 0.017 AUPRC while providing a speedup of $> 4\times$.

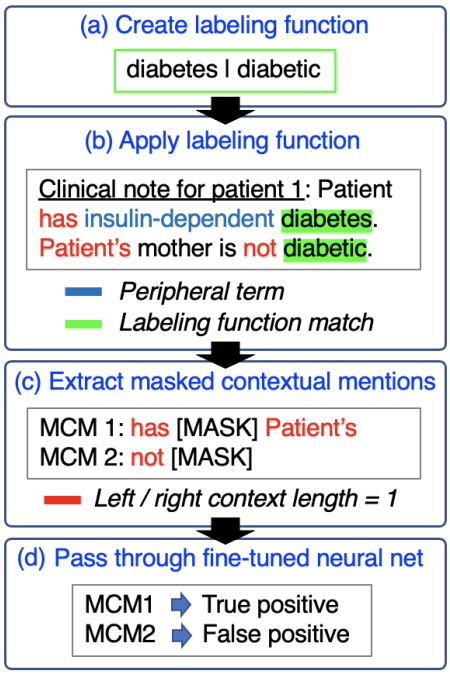

Figure 1: The workflow of HyDE. (a) a clinician develops a simple labeling function; (b) the labeling function is applied to the clinical note; (c) masked contextual mentions are extracted, including masked peripheral terms and contexts; (d) the masked contextual mentions are passed through the fine-tuned language model to identify false positives.

2 Methods

Our proposed method, HyDE (hybrid diagnosis extractor), aims to accurately identify the earliest occurrence of specific diseases in clinical patient encounter notes. We accomplish this by using a combination of labeling functions and a fine-tuned biomedical language model. The labeling functions are designed to be simple and identify as many mentions of the disease as possible, including false positives. The neural network is then used to differentiate between the true positives and false positives by analyzing small segments of the clinical notes around the location identified by the labeling functions. This approach allows for identifying potential mentions of the disease, while also utilizing the neural network to improve precision. It is worth noting that the components of HyDE are modular, allowing for the substitution of other methods for identifying disease-specific mentions beyond the labeling functions used in this paper. For example, Trove (Fries et al., 2021), offers ontology-based labeling functions that eliminate the need for coding task-specific labeling rules.

Our method (Fig. 1), involves the following steps: The user first develops a simple *labeling function* for the disease of interest. In the case of diabetes, this could be the regular expression `diabetes | diabetic`. This labeling function is then applied to the clinical notes to identify mentions of the disease. Additionally, the user identifies *peripheral terms* that frequently appear before or after mentions of the disease, such as `insulin-dependent` or `mellitus` in the case of diabetes. The text matching the labeling function and peripheral terms are then replaced with `[MASK]`, and a context around the resulting mask is extracted, resulting in a *masked contextual mention (MCM)*. These MCMs are used to fine-tune a biomedical language model to determine whether the context suggests that the patient actually has the condition in question. We hypothesize that this approach allows the language model to generalize to various conditions without additional training. Thus, for a zero-shot transfer to other diseases, only a simple disease-specific labeling function and peripheral terms are required. We adopt the term zero-shot in this context as each disease comes with distinct comorbidities, symptoms, and interventions.

2.1 Dataset

After obtaining approval from the institutional review board, we obtained ∼8.8 million clinical notes from 23,467 adult patients who had an encounter at our tertiary care center between 2012 and 2018.

2.2 Disease Phenotypes

We apply our electronic phenotyping method to five chronic diseases: hypertension (*HTN*), diabetes mellitus (*DM*), osteoporosis (*OST*), chronic kidney disease (*CKD*), and ischemic heart disease (*IHD*). These diseases were selected due to their high prevalence (HTN, 2021; DM, 2022; CKD, 2021; IHD, 2022; Clynes et al., 2020), the costs they incur to the healthcare system, and the potential for positive intervention (Blankemeier et al., 2022). For initial model training, we used hypertension as it is the most prevalent of these diseases (affecting 116 million in the US) (HTN, 2021) and we hypothesize that it generates the most diverse MCMs. Table 6 shows the labeling functions that we used to extract these mentions for each disease.

2.3 Data Labeling

Mask Contextual Mention Categories: We manually identified 6 categories of MCMs - (0) true positive; (1) false positive (otherwise unspecified); (2) referring to someone other than the patient; (3) referring to the patient but negated; (4) providing information / instructions / conditional statements (i.e. instructions for how to take a medication); (5) uncertain (i.e. differential diagnosis). Thus, category 0 is the true positive category and categories 1 - 5 are false positive categories. We formulate this problem as a binary classification where categories 1 - 5 are merged into class 1.

Amplifying False Positive Examples: The prevalence of false positives from our labeling functions were relatively low (Table 3). We thus sought to increase the number of category 2 false positive examples in our training dataset beyond the baseline prevalence of the 250 random MCM samples that were initially labeled (RS in Table 1). We applied a family labeling function to randomly sampled MCMs. This labeling function is positive if an MCM contains any term listed in A.1 relating to familial mentions. We generated 200 such category 2 amplified examples for subsequent labeling. Based on the annotations, we found that only 1.5% of the examples selected by this labeling function were actually true positives examples.

To increase the number of category 3 false positive examples, we applied the Negex algorithm (Chapman et al., 2001) to a separate set of randomly sampled masked contextual mentions. For further details see A.2. Based on manual annotation of 200 such examples, we found that 22% of the examples selected by this labeling function were actually true positive examples.

Filtering Masked Contextual Mentions: Applying the disease-specific labeling functions generated 827k, 555k, 87k, 199k, and 80k notes for HTN, DM, OST, CKD, and IHD respectively from roughly 8.1 million clinical notes (Table 4). Since clinical notes often contain duplicate information from multiple patient visits, we deduplicate the MCMs by comparing the 20 characters on either side of the masked mentions associated with a particular patient. If these characters are the same across multiple MCMs, we keep the MCM that was authored first and discard the others. Deduplication allows us to reduce the number of masked contextual mentions by $3.3\times$, $3.6\times$, $4.2\times$, $3.7\times$, and $3.3\times$ for HTN, DM, OST, CKD, and IHD respectively (Table 4). This method can be applied at inference to increase the computational efficiency of HyDE. Additionally, the length and number of MCMs per clinical note represents an average of 9% of the full notes for a context length of 64 words, which can improve the efficiency of inference on large datasets.

Active Learning: To further improve the performance of HyDE, we implement a human-in-the-loop uncertainty-based active learning strategy. This involves multiple iterations of training where after each iteration, 100 examples with corresponding probabilities closest to 0.5 are manually labeled and added to the training dataset for the next training iteration. Table 1 shows performance across the active learning iterations (A1-A4).

2.4 Model Training

We select PubMedBERT (Gu et al., 2021) (100 million parameters) as the model that we fine-tune due to its simple architecture and widespread validation. We use a train batch size of 8, an Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of 3e-5. We train for 25 epochs and choose the model checkpoint with the best validation set performance. 1,150 HTN examples are used for training and 250 HTN examples are used for validation. For disease specific fine-tuning experiments,

between 90 and 100 disease-specific examples are used for both validation and training. There was no overlap between the patients used for the hypertension training and validation sets and the patients used for test sets as well as disease-specific validation sets. Our test sets consisted of 442 - 500 labeled cases for each disease.

2.5 Evaluation

While labeling functions can be evaluated at a note level, we evaluate at a MCM-level since a single clinical note can consist of multiple MCMs. Furthermore, disease assignment based on clinical notes can be combined with assignment based on structured EHR, increasing the number of patients that are identified. Thus, we want to ensure high precision in identifying patients using clinical notes. For each MCM, we measure the fine-tuned language model's ability to correctly classify it as either true positive or false positive using area under the precision recall curve (AUPRC) and F1.

For our labeling function baseline (LF in Table 2), we use both the family labeling function described previously and Negex (Chapman et al., 2001). Although additional terms could be added to this labeling function, those same terms could also be added to HyDE, making this a fair comparison.

We also include a Word2Vec baseline in our comparison (Mikolov et al., 2013b,a). This technique leverages a pre-trained model which has been trained on a corpus of around 100 billion words from Google News. For each MCM, we aggregate word embeddings by calculating their mean and then train an XGBoost model (Chen and Guestrin, 2016) over the computed averages of the HTN training dataset MCM embeddings. To optimize the performance of our XGBoost model, we fine-tune its hyperparameters by conducting a grid search using our HTN validation dataset. It's worth mentioning that this strategy does not retain the sequential order of words.

To demonstrate the generalizability of our method on external data, we apply it to the assertion classification task from the 2010 i2b2/VA Workshop on Natural Language Processing (Uzuner et al., 2011). This dataset consists of 871 progress reports annotated with medical problems that are further classified as present, absent, possible, conditional, hypothetical, or not associated with the patient. We mapped the present category to class 0 and collated all other categories under class 1.

Table 1: Test set AUPRC comparison of the Word2Vec (W2V) baseline and fine-tuned PubMedBERT models (all rows except the first) using various training dataset compositions. Notation: RS - random MCM samples with baseline prevalence of false positive examples. C - additional category 2 amplified and category 3 amplified MCMs. A1, A2, A3, and A4 - additional MCMs labeled during four active learning iterations. SL - supervised learning. The test set sizes are 500, 455, 466, 442, 458 respectively for HTN, DM, OST, CKD, and IHD. * indicates that the W2V baseline was trained using the full RS+C+A4 training dataset.

Method	SL	Zero-Shot			
	HTN	DM	OST	CKD	IHD
W2V*	0.52	0.70	0.53	0.59	0.83
RS	0.60	0.73	0.59	0.71	0.82
RS+C	0.77	0.85	0.65	0.75	0.92
RS+C+A1	0.75	0.86	0.72	0.81	0.95
RS+C+A2	0.82	0.88	0.76	0.84	0.96
RS+C+A3	0.83	0.89	0.77	0.86	0.96
RS+C+A4	0.85	0.90	0.82	0.84	0.95

We used regular expressions to extract mentions of HTN, DM, OST, CKD, and IHD. We filtering out diseases with less than 30 mentions. Consequently, our external validation was conducted on HTN, DM, and CKD.

3 Results

Supervised and Zero-Shot Model Performance: Table 1 depicts AUPRC performance of our Word2Vec (W2V) baseline compared to fine-tuned PubMedBERT models trained with various training dataset compositions (all rows except the first). We demonstrate supervised performance on HTN, as well as zero-shot generalization to DM, OST, CKD, and IHD. The performance of HyDE surpasses that of our labeling function baseline by 44 points in F1 score and our Word2Vec baseline by 24 points in F1 score on average (Table 2). We find that fine-tuning the best PubMedBERT model (RS+C+A4 training dataset) on $\sim$100 additional disease-specific examples does not significantly improve performance, with scores of 0.91, 0.84, 0.81, and 0.95 on DM, OST, CKD, and IHD, respectively. This supports the conclusion that our model generalizes well to other diseases, without requiring disease-specific fine-tuning. On the external i2b2/VA dataset we achieve the following AUPRC scores without any additional finetuning - 0.79 for HTN (336 patients), 0.99 for DM (213 patients), and 0.95 for CKD (45

Table 2: F1 score comparison of the labeling function baseline (LF), the Word2Vec (W2V) baseline, and the RS+C+A4 fine-tuned PubMedBERT model. * indicates that the W2V baseline was trained using the full RS+C+A4 dataset.

Method	SL	Zero-Shot			
	HTN	DM	OST	CKD	IHD
LF	0.39	0.41	0.18	0.28	0.48
W2V*	0.41	0.61	0.48	0.54	0.68
RS+C+A4	0.74	0.81	0.75	0.74	0.89

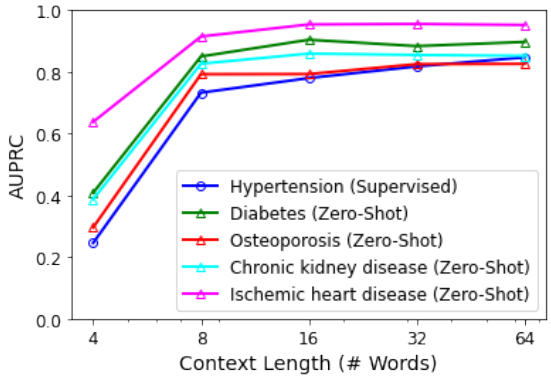

Figure 2: AUPRC of models using the RS+C+A4 data versus context length (words). Here context length is the number of words in the left context plus the number of words in the right context.

patients).

Context Length Ablation: Fig. 2 shows that RS+C+A4 (RS: 250 random MCM samples; C: 400 category 2 and 3 amplified MCMs; A4: 400 samples from active learning) trained models saturate with increasing context lengths. Table 5 shows that reducing the context length from 64 words to 16 words speeds up the model by 4.5x while only lowering average AUPRC by 0.017. From Table 4 we observe that this represents an average of 2.3% of the full clinical notes among notes that contain at least one MCM.

4 Conclusion

With its minimal setup, computational efficiency, and generalization capability, HyDE offers a promising tool for electronic phenotyping from unstructured clinical notes. By improving the ability to extract patient health status, we hope that HyDE will enable more informative large scale studies using EHR data, ultimately leading to public health insights and improved patient care.

5 Limitations

HyDE has yet to be tested in a large-scale and multi-site setting, which may offer more generalization challenges. Furthermore, an evaluation of note-level classification performance was not conducted. Although we expect that HyDE would perform well under such an evaluation, this would require heuristics to aggregate multiple MCMs per note.

6 Ethics Statement

The authors have carefully considered the implications of their work, including potential positive and negative impacts. A potential risk associated with this approach would be the leakage of protected health information (PHI) following a release of the model. To mitigate this risk, we will conduct a thorough review of the training data and consult with experts before deciding to release the model. Additionally, the authors have reviewed the ACM Code of Ethics and Professional Conduct document and attest that this work adheres to the principles outlined in that document.

References

2021. Chronic kidney disease in the united states, 2021.

2021. Facts about hypertension.

2022. Heart disease facts.

2022. National diabetes statistics report.

Juan M Banda, Martin Seneviratne, Tina Hernandez-Boussard, and Nigam H Shah. 2018. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science*, 1:53.

Louis Blankemeier, Isabel Gallegos, Juan Manuel Zambrano Chaves, David Maron, Alexander Sandhu, Fatima Rodriguez, Daniel Rubin, Bhavik Patel, Marc Willis, Robert Boutin, et al. 2022. Opportunistic incidence prediction of multiple chronic diseases from abdominal ct imaging using multi-task learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 309–318. Springer.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Juan M Zambrano Chaves, Akshay S Chaudhari, Andrew L Wentland, Arjun D Desai, Imon Banerjee, Robert D Boutin, David J Maron, Fatima Rodriguez,

Alexander T Sandhu, R Brooke Jeffrey, et al. 2021. Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach. *medRxiv*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Michael A Clynes, Nicholas C Harvey, Elizabeth M Curtis, Nicholas R Fuggle, Elaine M Dennison, and Cyrus Cooper. 2020. The epidemiology of osteoporosis. *British Medical Bulletin*.

Erik R Dubberke, Humaa A Nyazee, Deborah S Yokoe, Jeanmarie Mayer, Kurt B Stevenson, Julie E Mangino, Yosef M Khan, Victoria J Fraser, et al. 2012. Implementing automated surveillance for tracking clostridium difficile infection at multiple healthcare facilities. *Infection Control & Hospital Epidemiology*, 33(3):305–308.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):1–11.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

David C Kaelber, Wendy Foster, Jason Gilder, Thomas E Love, and Anil K Jain. 2012. Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. *Journal of the American Medical Informatics Association*, 19(6):965–972.

Elizabeth FO Kern, Miriam Maney, Donald R Miller, Chin-Lin Tseng, Anjali Tiwari, Mangala Rajan, David Aron, and Leonard Pogach. 2006. Failure of icd-9-cm codes to identify patients with comorbid chronic kidney disease in diabetes. *Health services research*, 41(2):564–580.

Matthew H Lee, Ryan Zea, John W Garrett, Peter M Graffy, Ronald M Summers, and Perry J Pickhardt. 2022. Abdominal ct body composition thresholds using automated ai tools for predicting 10-year adverse outcomes. *Radiology*, page 220574.

Paea LePendu, Srinivasan V Iyer, Cédrick Fairon, and Nigam H Shah. 2012. Annotation analysis for testing drug safety signals using unstructured clinical notes. In *Journal of biomedical semantics*, volume 3, pages 1–12. Springer.

Fernando Martin-Sanchez and Karin Verspoor. 2014. Big data in medicine is driving big changes. *Yearbook of medical informatics*, 23(01):14–20.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. 2021. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637.

Ethan Steinberg, Yizhe Xu, Jason Fries, and Nigam Shah. 2023. Self-supervised time-to-event modeling with structured medical records. *arXiv preprint arXiv:2301.03150*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Wei-Qi Wei, Cynthia L Leibson, Jeanine E Ransom, Abel N Kho, Pedro J Caraballo, High Seng Chai, Barbara P Yawn, Jennifer A Pacheco, and Christopher G Chute. 2012. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association*, 19(2):219–224.

Jingqing Zhang, Luis Bolanos Trujillo, Tong Li, Ashwani Tanwar, Guilherme Freire, Xian Yang, Julia Ive, Vibhor Gupta, and Yike Guo. 2021a. Self-supervised detection of contextual synonyms in a multi-class setting: Phenotype annotation use case. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8754–8769, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021b. Knowledge-rich self-supervision for biomedical entity linking. *arXiv preprint arXiv:2112.07887*.

A Appendix

A.1 Family Labeling Function

The family labeling function is positive if any of the following terms match within a masked con-

Table 3: Distribution of the different categories within 500 randomly sampled masked contextual mentions in percent. Cat denotes category. The categories are defined as follows. 0 - true positive; 1 - false positive (otherwise unspecified); 2 - referring to someone other than the patient; 3 - referring to the patient but negated; 4 - providing information / instructions / conditional statements (i.e. instructions for how to take a medication; "if you feel this way, do this"); 5 - uncertain (i.e. differential diagnosis; "likely"; "workup for").

Cat	Type	HTN	DM	OST	CKD	IHD
0	+	87.6	74.0	79.2	77.6	46.2
1	-	0.6	0.6	1.8	0.6	0.4
2	-	4.4	11.6	4.4	3.0	17.6
3	-	2.8	4.8	1.8	4.4	14.8
4	-	0.8	5.6	6.8	6.4	15.4
5	-	3.8	3.4	6.0	8.0	5.4

textual mention: `relative`, `relatives`, `family`, `father`, `mother`, `grandmother`, `grandfather`, `sister`, `brother`, `sibling`, `aunt`, `uncle`, `nephew`, `niece`, `son`, `daughter`, `cousin`, `parents`.

A.2 Negex Algorithm

In order to increase the recall of the Negex (Chapman et al., 2001) algorithm for manual labeling in order to amplify false positives for HyDE training, we modified it slightly to allow negative terms to match within 7 words of the mention, rather than 5. However, for the labeling function baseline we used Negex with a conventional window of 5 words, as opposed to the 7 word window used during HyDE training.

We modify the Negex keywords slightly based on manual examination of the MCMs. The original keywords were extracted from the negspaCy en_clinical termset. This function is positive if any of the following terms appear within the specified number of words before the disease mention: `declined`, `denied`, `denies`, `denying`, `no sign of`, `no signs of`, `not`, `not demonstrate`, `symptoms atypical`, `doubt`, `negative for`, `no`, `versus`, `without`, `doesn't`, `doesnt`, `don't`, `dont`, `didn't`, `didnt`, `wasn't`, `wasnt`, `weren't`, `werent`, `isn't`, `isnt'`, `aren't`, `arent`, `cannot`, `can't`, `cant`, `couldn't`, `couldnt'`, `never`, `none`, `resolved`, `absence of` or if any of the following terms appear within the specified number of words after the disease mention: `declined`, `unlikely`, `was not`, `were not`, `wasn't`, `wasnt`, `weren't`, `werent`, `not`, `no`, `none`.

A.3 Qualitative Evaluation of Active Learning Examples

Qualitatively, the examples surfaced during active learning appear to be challenging cases. For example, some were examples that would have been counted as false positives by Negex but shouldn't be. One such example is "Insulin dependent diabetes mellitus ¬ø [MASK] No past medical history pertinent negatives". Here, ¬ø denotes a de-identified date. Another challenging example is "4. Screening for [MASK]". Often when items are enumerated, they indicate a positive diagnosis. However, in this case, the patient was only screened for the condition.

Table 4: HyDE neural network computational efficiency. For reference, the average length of the 8.8 million clinical notes in the dataset is 375 words. We filter these 8.8 million notes down to 8.1 million notes by note type. We include the most common note types in our dataset: "Progress Note", "Inpatient", "ED Note", "Consultation Note", "Letter", "Other Note", "Nursing Sign Out Note", "History and Physical", "Outpatient", "IP Consult", and "Discharge/Transfer Summary". The number of MCMs generated after deduplication based on local context of 20 characters is shown below. These numbers vary depending on the exact form of the labeling functions used.

Metric	HTN	DM	OST	CKD	IHD
Before deduplication					
Number of notes with MCMs	827k	555k	87k	199k	80k
Number of MCMs	1,616k	1,264k	127k	449k	125k
MCMs per note	2.0	2.3	1.5	2.3	1.6
Average size of notes with MCMs (words)	1,256	1,247	1,508	1,374	1,473
% notes represented by MCMs (64 word context)	10%	12%	6%	11%	7%
Number of MCMs after deduplication	495k	353k	30k	120k	38k
MCM reduction through deduplication	3.3x	3.6x	4.2x	3.7x	3.3x

Table 5: Inference time versus context length. All experiments are performed on a single 12GB Titan Xp GPU. Analysis is done using 15,000 MCMs and the times reported are the total time spent for each task while processing the 15,000 MCMs. Batchsizes are increased in increments of 100 until they no longer fit on the GPU.

Context length (words)	16	32	64
Batch size (MCMs)	3800	2000	1000
Total inference time (s)	17.47	43.33	79.07
Data transfer CPU to GPU time (s)	14.92	39.23	72.08
Tokenization time (s)	0.81	1.10	1.63
Model run time (s)	0.44	1.58	4.09
MCMs / second	859	346	190

Table 6: Labeling functions used to extract masked contextual mentions. HTN, DM, OST, CKD, and IHD stand for hypertension, diabetes, osteoporosis, chronic kidney disease, and ischemic heart disease respectively.

Disease	Labeling Function
HTN	(\s+hypertension)|(\s+HTN)
DM	(\s+diabetes)|(\s+DM2)|(\s+DM\s+)|(\s+T2DM)
OST	(\s+osteoporosis\s+)|(\s+osteoporotic\s+)
CKD	(\s+kidney failure)|(\s+nephropathy)|(\s+CKD\s+)|(\s+kidney disease)| (\s+chronic kidney disease)|(\s+renal disease)|(\s+ESRD\s+)
IHD	(\s+NSTEMI\s+)|(\s+myocardial ischemia)|(\s+ischemic heart disease)| (\s+cardiac ischemia)|(\s+myocardial infarction)|(\s+myocardial necrosis)| (\s+coronary heart disease)|(\s+coronary artery disease)|(\s+heart attack)

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 5

☒ A2. Did you discuss any potential risks of your work?
Section 6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
End of the abstract and end of the introduction (section 1)

☑ A4. Have you used AI writing assistants when working on this paper?
We used ChatGPT to propose suggestions for improving the grammar and phrasing of author generated writing. We used this parts of each section of the paper, but did not always use the suggestions generated by ChatGPT and we always modified the suggestions.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 2 (methods) and section 3 (results)

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 2.5 (model training) and Table 5 in the appendix.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 2.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3 (results)

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2.3 (data labeling)

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 2.3 (masked contextual mention categories)

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Annotation was done by the authors of the paper.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Annotation was done by the authors of the paper.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 2.1 (dataset)

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Annotation was done by the authors of the paper.