A Taxonomy for Advancing Systematic Error Analysis in Multi-Site Electronic Health Record-Based Clinical Concept Extraction

A taxonomy for advancing systematic error analysis in multi-site electronic health record-based clinical concept extraction

Article *in* Journal of the American Medical Informatics Association · May 2024

DOI: 10.1093/jamia/ocae101

CITATIONS
2

READS
148

14 authors, including:

Sunyang Fu
The University of Texas Health Science Center at Houston
141 PUBLICATIONS **1,552** CITATIONS
SEE PROFILE

Liwei Wang
Mayo Clinic
141 PUBLICATIONS **3,439** CITATIONS
SEE PROFILE

Andrew Wen
Mayo Clinic - Rochester
87 PUBLICATIONS **1,079** CITATIONS
SEE PROFILE

Nansu Zong
University of California, San Diego
69 PUBLICATIONS **875** CITATIONS
SEE PROFILE

A taxonomy for advancing systematic error analysis in multi-site electronic health record-based clinical concept extraction

Authors:
Sunyang Fu, Ph.D., MHI[12], Liwei Wang, Ph.D.[12], Huan He, Ph.D.[6], Andrew Wen, M.S.[2], Nansu Zong, PhD[1], Anamika Kumari, MS[3], Feifan Liu, Ph.D.[3], Sicheng Zhou[5], Rui Zhang, Ph.D.[5], Chenyu Li[4], Yanshan Wang, Ph.D.[4], Jennifer St Sauver Ph.D.[1], Hongfang Liu, Ph.D.[12], Sunghwan Sohn, Ph.D.[1]

Affiliations:
[1]Mayo Clinic, Rochester, Minnesota
[2]University of Texas Health Science Center at Houston, Houston, Texas
[3]University of Massachusetts Chan Medical School, Boston, Massachusetts
[4]University of Pittsburgh, Pittsburgh, Pennsylvania
[5]University of Minnesota, Minneapolis, Minnesota
[6]Yale University, New Haven, Connecticut

Corresponding author:
Dr. Sunghwan Sohn, Department of AI and Informatics, Mayo Clinic, 200 First St SW, Rochester, MN 55905 (email: sohn.sunghwan@mayo.edu)

Keywords: error analysis; natural language processing; electronic health record.

Abstract

Background: Error analysis plays a crucial role in clinical concept extraction, a fundamental subtask within clinical natural language processing (NLP). The process typically involves a manual review of error types, such as contextual and linguistic factors contributing to their occurrence, and the identification of underlying causes to refine the NLP model and improve its performance. Conducting error analysis can be complex, requiring a combination of NLP expertise and domain-specific knowledge. Due to the high heterogeneity of electronic health record (EHR) settings across different institutions, challenges may arise when attempting to standardize and reproduce the error analysis process.

Objectives: This study aims to facilitate a collaborative effort to establish common definitions and taxonomies for capturing diverse error types, fostering community consensus on error analysis for clinical concept extraction tasks.

Materials and Methods: We iteratively developed and evaluated an error taxonomy based on existing literature, standards, real-world data, multisite case evaluations, and community feedback. The finalized taxonomy was released in both .dtd and .owl formats at the Open Health Natural Language Processing Consortium. The taxonomy is compatible with several different open-source annotation tools, including MAE,
Brat, and MedTator.

Results: The resulting error taxonomy comprises 43 distinct error classes, organized into 6 error dimensions and 4 properties, including model type (symbolic and statistical machine learning), evaluation subject (model and human), evaluation level (patient, document, sentence, and concept), and annotation examples. Internal and external evaluations revealed strong variations in error types across methodological approaches, tasks, and EHR settings. Key points emerged from community feedback, including the need to enhancing clarity, generalizability, and usability of the taxonomy, along with dissemination strategies.

Conclusion: The proposed taxonomy can facilitate the acceleration and standardization of the error analysis process in multi-site settings, thus improving the provenance, interpretability, and portability of NLP models. Future researchers could explore the potential direction of developing automated or semi-automated methods to assist in the classification and standardization of error analysis.

Introduction.

The digital transformation of healthcare has revolutionized the documentation and management of electronic health records (EHRs), thereby increasing the need for innovative informatics solutions to facilitate the secondary use of EHRs for healthcare delivery and clinical research. However, a substantial amount of valuable information in EHRs is written in free text. To transform routinely generated EHR data into actionable knowledge, scalable approaches are needed. Natural language processing (NLP) offers computational approaches to unlock clinical information in high throughput, which is essential to the advancement of the digitization of healthcare. Particularly, clinical concept extraction is one of the most common clinical NLP sub-tasks and refers to the automated extraction of predetermined clinical concepts from unstructured textual data such as clinical notes, radiology reports, operative reports, and pathology reports [1]. This process involves identifying instances of clinical concepts (known as concept mention detection or named entity recognition) and encoding them in a structured format for further analysis (i.e., concept encoding) [2].

Error analysis is a process used to document errors that appear after an NLP model run. Analysis is usually accomplished by manually reviewing the types of errors (e.g., contextual or linguistic contributors to the error causes), identifying the appropriate rationale, followed by adjusting the model to fix errors (i.e., refinement) to achieve optimal performance. Error analysis can be performed during the model development or training phase to help improve the model's performance and after the model has been evaluated on a blinded test set to provide additional insights into the model's explainability and limitations. Another common scenario for error analysis is when a model faces portability issues. Multi-institutional portability measures the out-of-the-box and post-refinement performance of a model when deploying it to EHRs from different institutions. For example, previous studies have found between 4% to 72% f1-score degradation in various domains, such as mental health[3], cardiology[4], and orthopedics[5]. Detailed error analysis can systematically study the variation among the clinical language and infer various reasons for heterogeneity factors such as different EHR systems, data quality, clinical practice, and documentation patterns across different institutions. Recognizing the need for external validation and representativeness of EHR data, there has been an increasing number of community efforts, such as the ENACT (Evolve to Next-Gen Accrual to Clinical Trials) NLP and OHDSI

NLP working group, aimed at improving the robustness and external validity of clinical NLP models.

In the past, researchers have designed and developed user interaction tools to improve the validity, reproducibility, and usability of the error analysis process. For example, Wu et al. developed Errudite, an interactive tool with data grouping functionality to improve the reproducibility and usability of error analysis [6]. iSEA is another error analysis pipeline, which automatically discovers semantically-grounded subpopulations with high error rates in the context of a human-in-the-loop interactive system [7]. However, in the context of clinical NLP, the process of conducting error analysis can be iterative and complex, requiring both NLP expertise and problem-specific domain knowledge. Compared with corpus annotation, which is a standardized process for gold standard creation, error analysis requires the evaluator to have a solid understanding of both medical and NLP tasks. Therefore, establishing common consensus is crucial, especially in the context of different clinical NLP tasks, because various heterogeneous factors can have a significant impact on the form and format of clinical text. Currently, there is still a lack of a unified effort to develop common definitions and taxonomy for capturing various heterogeneous error types. Our primary goal for this study was to establish community consensus on error analysis for clinical NLP tasks to enhance NLP model explainability and portability in multisite studies. Here, we propose error taxonomy for systematic error analysis in the context of EHR-based studies.

Materials and methods.

The development of the error taxonomy involved an iterative process that included three phases (development, evaluation, and dissemination). For error taxonomy development, the process of gathering existing resources to create the initial prototype version. At the same time, we deployed two previously validated NLP models to two additional hospital EHR sites to gather real-world errors. An iterative taxonomy development process was conducted to create the first version of the error taxonomy based on the errors from two NLP models (falls and delirium identification) developed by Mayo Clinic and refined on a large corpus to expand coverage. The proposed taxonomy was then evaluated in external real-world studies on extracting breast cancer, transportation, and social determinants of health (SDoH) information at three different institutions. A literature review was conducted to assess the coverage of the taxonomy in relation to more recently published articles, including studies that focus on large language models (LLM) and

prompt-based concept extractions. We also collect community feedback from NLP working groups and AMIA (American Medical Informatics Association). Lastly, we disseminate the taxonomy based on community standards and present a case study on the MedTator visual error analytics module. The overview of the development, evaluation, and dissemination process is presented in Figure 1.

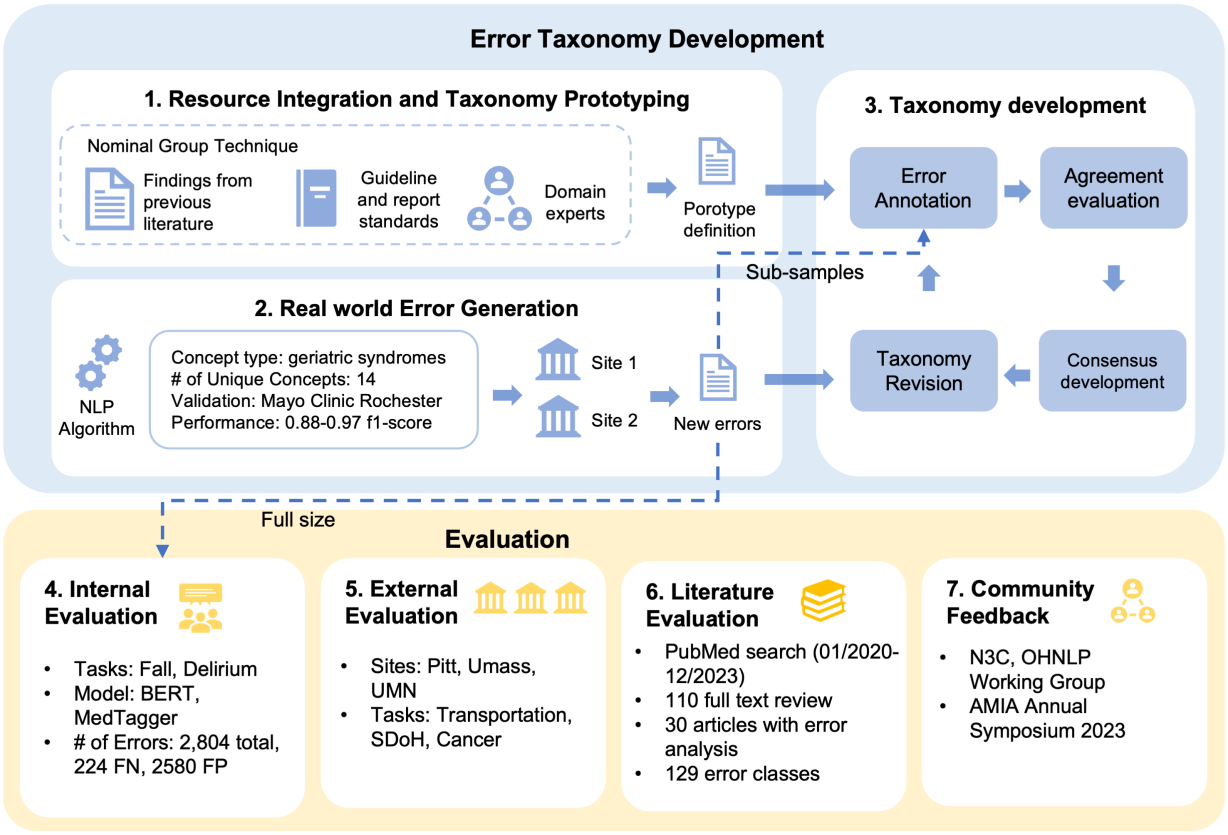

Figure 1. Overview of Error Taxonomy Development, Evaluation, and Dissemination Process. Site 1: a network of community-based health systems, Site 2: a nonprofit community-based hospital, Pitt: The University of Pittsburgh, Umass: University of Massachusetts Chan Medical School, UMN: The University of Minnesota.

Resource integration taxonomy and prototyping. The development of the proposed taxonomy was designed based on findings from our previous survey paper[2]. Specifically, we focused on 19 articles with comprehensive reporting on NLP features[8-22]. The first author (S.F.) prototyped

the initial definition, which was then iteratively reviewed and revised by S.S., L.W., and A.W. in word choice, grouping, and definition. The initial definition includes 17 unique error classes.

Real-world error generation. To collect real-world NLP errors, we deployed and ran two NLP models (NLP-CAM for delirium-related concept detection[23] and NLP-Fall[24] for falls detection), which were originally developed at Mayo Clinic Rochester and refined on EHRs from two additional health care systems participating in the Rochester Epidemiology Project: (Site 1) a network of community-based health systems and a nonprofit community-based clinic and hospital (Site 2).

- The NLP-CAM model is a rule-based model based on the MedTaggerIE framework, containing 15 unique delirium-related neuropsychological concepts such as agitation, confusion, and hallucination which are part of the Confusion Assessment Method (CAM) rubric for delirium identification (NLP-CAM). The model achieved an f1-score of 0.96 among 300 patients[23].
- The NLP-Fall model is a BERT model that identifies fall occurrences from clinical notes. The model achieved an f1-score of 0.97 among 64,312 test sentences in the original Mayo Clinic Rochester setting[24].

To establish a gold standard dataset, we first conducted corpus annotation at Site 1 and Site 2. We selected two case studies (delirium and falls) that had undergone a rigorous chart review process [24, 25]. The workflow was guided by the TRUST process [26, 27], which has been adopted by the National Center for Data to Health (CD2H) as best practice. Briefly, the process involves cohort development, data retrieval, quality control iteration (e.g., annotation training, double annotation, IAA measurement) consensus development, and adjudication. All case studies followed a standardized annotation guideline [24, 25, 28, 29] (Supplementary Appendix 5). For the delirium study, we annotated 354 Site 1 patients and 193 Site 2 patients. For the fall study, we performed annotation on 542 Site 1 patients and 372 Site 2 patients. The annotation process was conducted by the same annotation team using the same annotation guidelines and processes[23, 24]. After the annotation, we ran two models on two institutions and measured the concept-level performance against human readers. Based on the deployment evaluation at concept level, NLP-CAM yielded out-of-box precision, recall, and f1-score of 0.355, 0.831, and 0.498 among 1,896 documents from Institution 1 and 0.478, 0.789, and 0.596 among 396 documents from Institution 2, respectively. NLP-Fall achieved out-of-box precision, recall, and f1-score of 0.117, 0.871, and

0.206 among 88354 sentences from Institution 1 and 0.757, 0.896, and 0.821 among 14807 sentences from Institution 2, respectively. A total of 2,804 unique errors were collected, including 224 false negatives and 2,580 false positives. A detailed breakdown of the error statistics and interpretation for performance degradation across sites can be found in Appendix 2.

Taxonomy development. The consensus development process consisted of three steps: (a) applying the prototyped error definition to real-world errors through a formal annotation process performed by two experienced informaticians (S.F. and L.W., both with 8+ years of experience in gold standard development and NLP methodology), supervised by another senior investigator (S.S.), (b) iteratively measuring agreement, organizing scientific meetings, and updating definitions until all issues were resolved, and (c) finalizing the error definition and adjudicating all disagreements once an acceptable agreement was reached (f1-score>0.7). The data used in the initial taxonomy development include two batches of samples (n=100) from previously collected errors from Site 1 and 2 EHR data. The agreements for two rounds of consensus development were 0.71 and 0.75 in f1-score. The top disagreed categories were "Exclusion" (20%), "Missing_Annotation" (18%), "Implied_Inference" (16%), "Certainty" (10%), and "Absence_of_Context" (9%). All disagreements were resolved through three rounds of consensus meetings. Modifications to the definition included distinguishing human and model errors, clarifying definitions (e.g., implication and synonym errors), and improving terminology clarity. Detailed guidelines for taxonomy development can be found in Appendix 5.

Internal evaluation. To assess whether the proposed taxonomy had comprehensive and in-depth coverage, the same informaticians applied the definition to a total of 2,804 unique errors (from Figure 1, Step 4). Rather than focusing on the consensus between two reviewers, this round of evaluation aimed to assess the coverage of the initial version of the taxonomy on a large scale. In total, we conducted two rounds of evaluation over an 8-month period. Both rounds involved manual reviews of all generated errors. The first round occurred after the taxonomy development phase and aimed to assess the coverage of the initial definition. The second round took place after the multisite evaluation and the collection of community feedback, with the goal of evaluating the frequency distribution differences among error classes.

External evaluation. The usability, coverage, and generalizability of the proposed taxonomy were further evaluated in three real-world case studies, including breast cancer, transportation, and social determinants of health (SDoH), from three institutions: The University of Minnesota

(UMN), The University of Pittsburg, and The University of Massachusetts Chan Medical School. Three unique NLP models were included in the evaluation including the CancerBERT[28], the nlp4transportation, and the SDOH-NER model (unpublished). The CancerBERT model was pre-trained and fine-tuned on the breast cancer corpus collected from the UMN Clinical Data Repository to extract breast cancer-related concepts. The nlp4transportation model, engineered for the University of Pittsburgh Medical Center Adult Spina Bifida Clinic, aims to identify patterns indicative of patient transportation usage. The SDOH-NER model was trained using the N2C2 Track 2 Challenge MIMIC Dataset to extract SDOH entities from clinical texts. The entire evaluation process lasts about six months. The initial weekly meeting was organized for the first three months. An evaluation guideline was developed to capture the meta and standardize the evaluation process across sites. Prior to the evaluation, basic data elements including task name, study area (name of disease, drug, or SDoH), Knowledge and level of expertise in clinical NLP (High [more than 5 years of experience]; Intermediate [2-4 years of experience]; Low [1-2 years of experience]), information about NLP methodology, model, and framework, information about annotation were collected. Prior to the evaluation, we scheduled two training sessions and performed a pilot evaluation. During the pilot evaluation, each site performed trial annotation on 10 samples. A weekly consensus meeting was scheduled for the first month. During the formal evaluation phase, each site randomly samples ~50 NLP errors (preferably 50% FP, 50% FN) from a real-world EHR-based clinical concept extraction task (e.g., disease, drug, and SDoH). EHR-based clinical concept extraction task: the task of automatically extracting pre-defined clinical concepts from unstructured EHR data. The study does not include relation extractor and coreference resolution. Additional details about each case study and evaluation process can be found in the Appendix.

Literature evaluation. The taxonomy was prototyped based on findings from our previous review study, which analyzed studies published between January 2009 and June 2019[2]. To evaluate the taxonomy's coverage to more recent publications, including those focusing on large language models (LLMs) and prompt-based concept extractions, we conducted a literature review evaluation study. We performed a PubMed search of clinical NLP articles published from January 2020 to December 2023 (with no overlap with our previous review study[2]). The initial search returned a total of 197 articles. After title and abstract screening, 110 remained eligible and proceeded to full-text review. Among them, we identified 30 articles with error analysis

descriptions. We manually collect error class (category), description, and examples provided by each study. We then assess the coverage of taxonomy by manually mapping each error class into exact match, partial match, unmatch, and irrelevant.

Community feedback collection. To collect community feedback on the design, content, and dissemination strategy of the taxonomy, we organized two phases of community engagement events. In phase one, we presented the taxonomy during N3C NLP working group meetings hosted by the Open Health Natural Language Processing (OHNLP) Consortium[30]. The OHNLP Consortium-funded NCATS Innovation Award is a community-based collaborative effort for promoting technology innovations and consortium-wide dissemination in clinical NLP. The taxonomy was presented in front of 15 NLP researchers and community leaders from more than 8 different institutions in two sessions. The MedTator annotation tool was used to facilitate live demonstration. The initial community engagement aims to gain feedback on the needs, preliminary studies, and feasibility of the task. In phase two, we presented our findings at the American Medical Informatics Association (AMIA) 2023 Annual Symposium to collect broader feedback from the community. Feedback was analyzed and organized into a high-level summary.

Results.

Error taxonomy. The derived error taxonomy (as of March 23, 2024) consists of 43 unique error classes grouped by six error dimensions: "Logic_Error", "Linguistic_Error", "Contextual_Error", "Annotation_Error", and "Other_Error" (Figure 2). Each error has four properties including model type (symbolic and statistical machine learning), evaluation subject (model and human), evaluation level (patient, document, sentence, and concept), and annotation examples. Evaluation level: evaluation is usually performed at one of several levels: a patient level, a document level, a concept level, multiple levels, or a defined episode level (with temporality). The specific level selected with which evaluation was performed was typically determined based on the specific task or application. For example, patient-level detection may be sufficient if the task is to detect patients with a disease or a disease phenotype. On the other hand, identification of the time of presentation likely requires an evaluation with a finer level of granularity.

To ensure the interoperability and usability of the taxonomy, we have released it in .dtd and .owl format at the OHNLP consortium. Detailed definitions for each error class and properties can be accessed at https://github.com/OHNLP/ErrorAnalysis. The taxonomy is also compatible with

several different open-source annotation tools such as MAE, Brat, and MedTator. We incorporated the taxonomy with an analytical module within the OHNLP-MedTator, a serverless text annotation tool to reuse the in-browser modules for corpus text processing [2] (https://ohnlp.github.io/MedTator/).

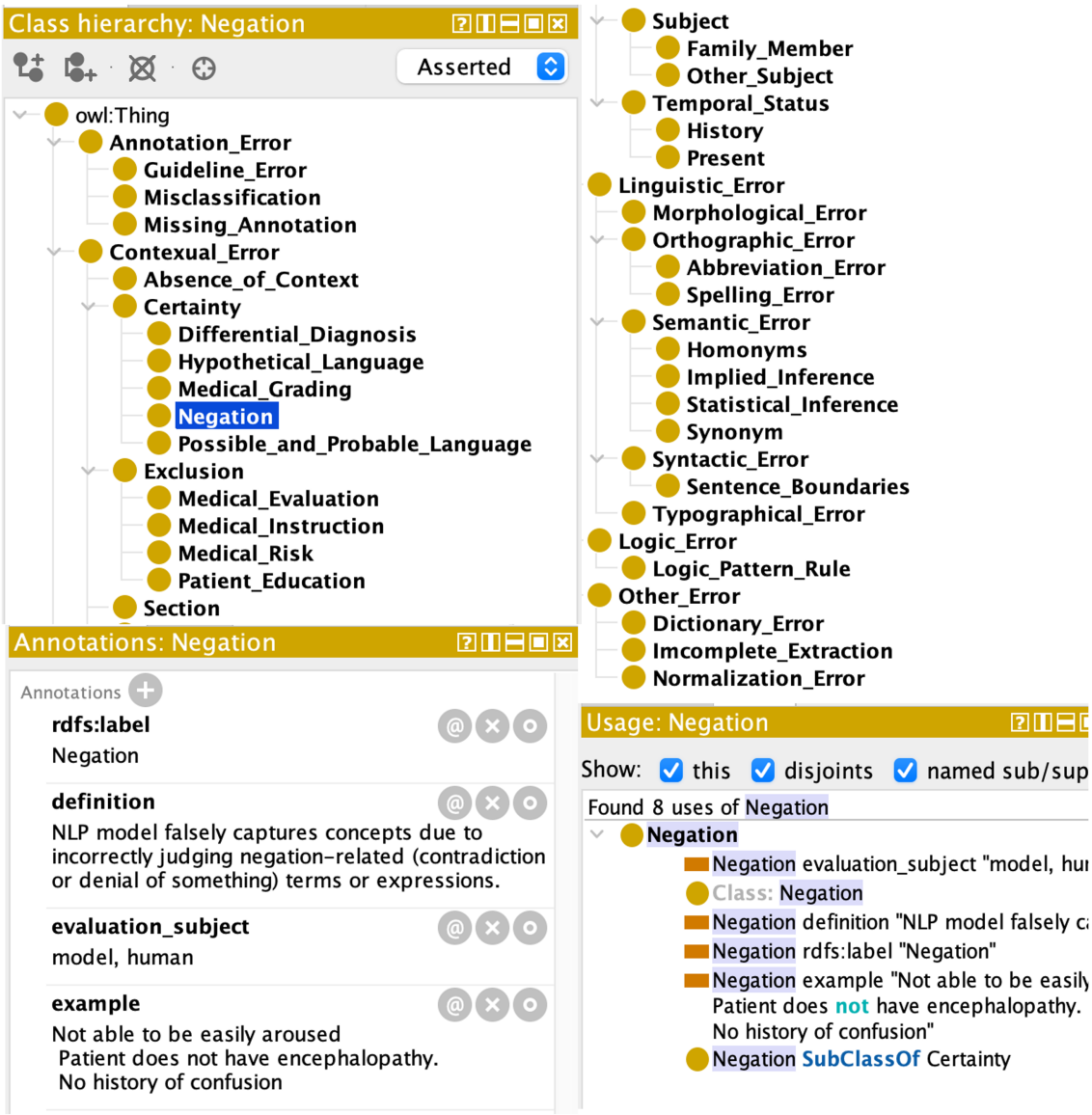

Figure 2. Overview of Error Taxonomy in OWL Format.

Internal evaluation. The evaluation of the large annotation corpus helped identify three additional sub-class categories, including "Patient_Education," "Risk_Assessment," "Medical_Evaluation,"

and "Medical_Risk" under "Exclusion" and "Differential_Diagnosis" under "Certainty." According to the second round evaluation, we found "Missing_Annotation" is the most common error type, occurring 1,634 times, making up 58.27% of all errors. "Negation" is the second most frequent error, with 608 occurrences, accounting for 21.68% of the total. Several other error types have lower frequencies, ranging from "Implied_Inference" and "Exclusion" at 4.10% and 3.28%, respectively, down to "Homonym" at just 0.04%. The detailed taxonomy coverage for all 2,804 errors can be found in Appendix 1. We further compared the class distribution between false positive and false negative cases (Figure 3). For false positive cases, two commonly discovered error dimensions were 'Annotation_Error' and 'Contextual_Error,' which included 'Missing_Annotation' and 'Negation.' Additionally, 'Exclusion' from the 'Contextual_Error' dimension was also observed. For false negative cases, the top error classes were 'Implied_Inference,' 'Morphological,' 'Synonym,' and 'Orthographic,' all of which fell under the 'Linguistic_Error' dimension. Overall, there was a substantial difference in error classes between false positive and false negative cases.

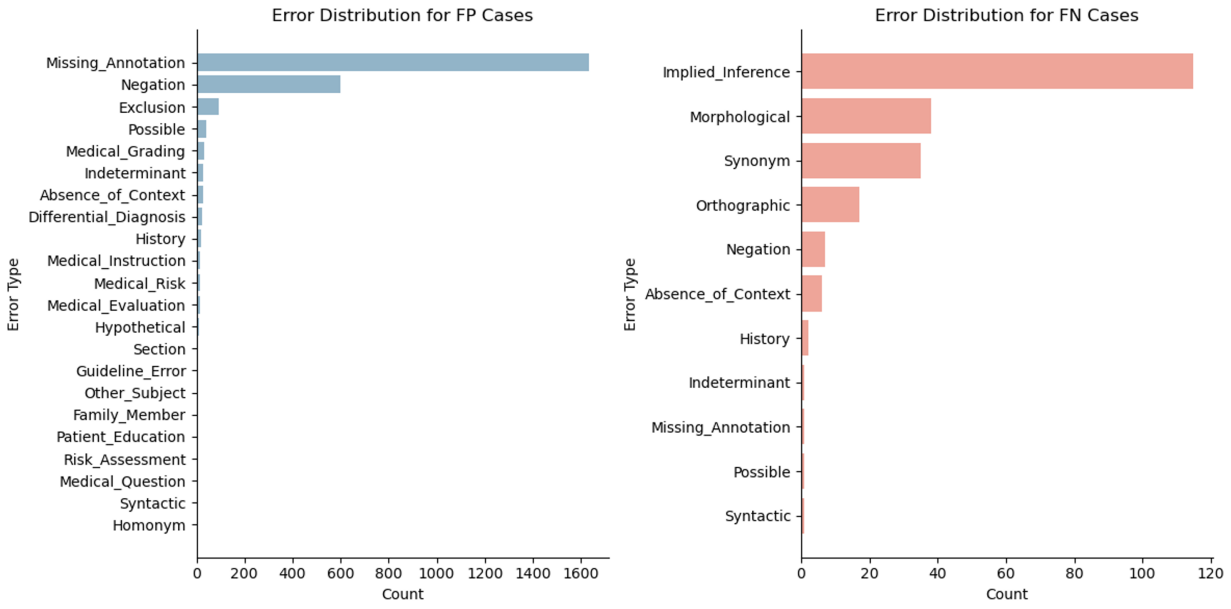

Figure 3. Comparison of Error Type Distribution between False Positive (Left) and False Negative (Right) Cases in Large Corpus Evaluation.

External evaluation. Figure 4 provides a detailed breakdown of different error class coverage across three case studies. In general, "Missing_Annotation", "Exclusion", and

"Absence_of_Context" were the most common error classes discovered. In the breast cancer study, the most frequently occurring error types included "Missing_Annotations", "Indeterminacy", and "Absence_of_Context". In the transportation study, common error types included "Missing_Annotations", "Exclusion", and "Status" and for the SDoH study, "Misclassification", "Absence_of_Context", as well as "Exclusion". The evaluation also helps identify two additional error classes including "Statistical_Inference" and "Incomplete_Extraction". Common examples include rare expressions although occur in training data, the model is still unable to fully capture all possible cases. In the breast cancer study, we encountered a limited number of cases in the 'Indeterminant' category for the CancerBERT classifier. These results highlight the varying levels of explainability between symbolic and statistical machine learning approaches.

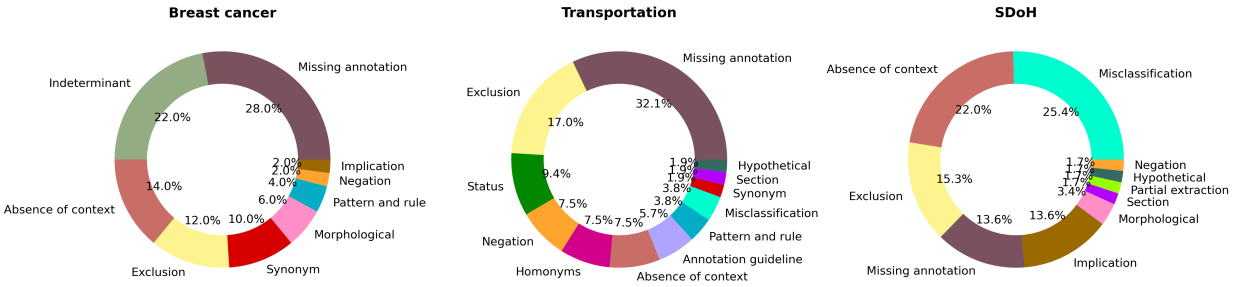

Figure 4. Visual Comparison of Error Class Across Three Differences

Literature evaluation. Among the 30 studies with error analysis descriptions, a total of 129 error classes were identified. Out of these 129 error classes, 86 were included in the mapping analysis, while 43 error classes were excluded. Specifically, we did not consider nine cases that were generically described as 'False Positive' or 'False Negative'. Ten cases were attributed to issues related to text data quality or research quality and were also not considered. While text data quality is relevant to NLP errors, we have confined the scope of our taxonomy to clinical concept extraction. Future efforts could explore the integration of our work with existing data quality-related taxonomies and ontologies. For example, errors due to missing clinical notes, miscommunication within clinical documents, and text content are corrected over time. We also did not consider using disease or concept names to categorize errors because this representation can be automatically summarized by each NLP task and may provide limited actionable

information for addressing the errors. For example, 'Medication Error' and 'Trauma Errors'. In addition, a few errors (<2%) lacked appropriate definitions for us to accurately determine which error class or subclass they belong to. Examples include 'Type Error', 'Colloquial Language', 'Domain Mismatch Errors', 'Complex Sentence Structure', and 'Intersentential'. Lastly, we removed nine cases related to errors in text generation tasks from LLMs. These cases were used to classify the harmfulness of the generated context rather than focusing on concept extraction tasks. For example, “Error reached patient did not cause harm”, “Error required intervention to sustain life” and “Error reached the patient and required monitoring”.

Out of the 86 error classes included in the mapping analysis, 72 (83%) were successfully mapped. Four unmapped error classes were added after the mapping analysis: 'Typographical Error' (10%), 'Dictionary Error' (3%), 'Normalization Error' (1%), and 'Relation Error' (1%).

Community feedback. Several key points emerged from the feedback, including the need for further disambiguation of error definitions to enhance clarity and provide concrete examples for each error class, particularly for challenging distinctions such as 'Certainty' versus Implied_Inference.' To enhance the generalizability of the taxonomy, there was a call for a more detailed evaluation of the results and their implications. The community also emphasized the importance of discussing the challenges in consensus development and proposing strategies for disseminating and advocating for the framework's adoption among end-users in broader practice. The majority of the feedback points were respectively considered and incorporated into the current version of the taxonomy. The feedback points that couldn't be addressed immediately are highlighted in the 'Future Study' section.

Visualization of error analysis. An example of an error annotation analytical module on MedTator is illustrated in Figure 5. The module provides post-error summary statistics (under the Error Summary Tab), Error Distribution, t-SNE visualization of error textual cosine similarity (Tag Distribution), and Error List. The visual analytics module helps to break down the number of false positive and negative cases, followed by error dimensions (annotation, contextual, linguistic, and logic), followed by each unique error type and specific concept. In addition to the analytical module, the error taxonomy can be encoded as an annotation schema to be further reused in other error analysis projects.

Figure 5. Visualization of Error Analytics based on MedTator

Discussion.

NLP models for EHR-based concept extraction require error analysis and refinement to perform well on EHRs from other institutions. This process often demands significant effort from NLP and domain experts. In our study, we proposed an error taxonomy aimed at accelerating and standardizing the error analysis process in multi-site settings with the goal of improving the provenance, interpretability, and portability of NLP models. Here, we discuss several aspects related to the quality and usability of error analysis, as well as the challenges encountered in the study.

Task Heterogeneity and Standardization. In the context of multisite EHR studies, variations in EHR system implementations, ETL (extract, transform, and load) processes, care and documentation practices, and patient populations can impact the representation, meaning, and definitions of text data across multiple institutions. Such variability becomes a huge barrier to the robustness of NLP systems. Based on the Kruskal-Wallis test on the error taxonomy distributions for each task group, we observed a significant difference among the groups ($p=0.01$). The most common error classes across three corpora were “Absence_of_context”, “Exclusion”,

"Missing_annotation", and "Negation". These four classes also yield high frequency in the internal evaluation under the delirium and fall tasks. We also observed that each corpus presented a few unique error types, including "Annotation_guideline" and "Homonyms" for Transportation, "Indeterminate" for the breast cancer study and "Partial_extraction" for the SDoH study. Further analysis indicated that such variability could be attributed to the complexity of the task, characteristics of clinical notes, and NLP approaches such as symbolic versus machine learning methods. In the breast cancer study, the model was designed to extract the breast cancer phenotypes of patients from clinical texts, i.e., Hormone receptor type, Hormone receptor status, Tumor size, Tumor site, Cancer grade, Histological type, Tumor laterality, and Cancer stage.

The various types of errors observed across different tasks and domains may stem from several factors, including the inherent nature of the tasks (such as varying definitions, levels of difficulty, and distributions of error-prone cases), as well as the techniques utilized. Certain NLP techniques exhibit specific weaknesses, leading to particular errors; for example, rules-based methods may struggle to capture contextual meaning, while machine learning approaches may produce incorrect learning outcomes when trained on limited samples for specific cases. This variability, arising from the heterogeneity of the EHR environment[31] and differing NLP development and evaluation methodologies[27], can present challenges in reproducibility. Our recent review study has also identified significant variations in reporting practices across NLP-assisted clinical studies[32]. These findings underscore the importance of establishing a taxonomy for systematic error analysis in order to address these issues effectively.

The proposed error taxonomy aims to capture various forms and formats of real-world NLP errors and standardize the error analysis process to provide a unified view of how different NLP models interact with heterogeneous EHR data. Through refinement and external evaluation, the taxonomy effectively captured the majority of relevant error cases. During the refinement process on the large corpus, we found a total of 31 indeterminate cases (1.11%), defined as cases that did not belong to any error classes. For external evaluation, we identified 11 cases (22%) for breast cancer, while there were no cases found for transportation or SDoH studies, respectively. These indeterminate cases were due to the limitation in the original formulation of the study. For example, the breast cancer study was designed as a sentence classification task using a domain-specific language model (CancerBERT). During the error analysis, we were only able to access sentence-level information due to data privacy constraints across institutions. When designing an

NLP measurement study[33], it is important to leverage additional information, such as other clinical documents, for comprehensive error analysis, even if the model does not require this information for decision-making. For the statistical machine learning NLP model, these indeterminate cases were primary due to missing relevant context needed to determine the behavior of the CancerBERT. For the symbolic MedTagger model, the majority of these indeterminate cases were due to the inability to associate the specific NLP rule with the prediction label. In other words, no pre-defined dictionary or keyword was found in the extracted mentions. This issue may be caused by the sentence segmentation pipeline of the NLP system, which will be further investigated in our future study.

Reproducibility and interpretability. In addition to standardization, another motivation for our study was to improve the provenance and interpretability of NLP results in the settings of multisite EHR data. The inherent latent variability in the documentation and textual information across different EHR environments can lead to potential uncertainties in the NLP model's behavior. Therefore, it is important to systematically capture the metadata of NLP information. In our prior work, we provided recommendations for the minimal reporting data elements for the NLP methodology and evaluation process[27, 32]. Using the proposed error taxonomy, we can further capture information related to how NLP interacts with the local EHR context. For example, in addition to reporting evaluation metrics, capturing statistics and examples of error cases can provide additional explanations for certain under-performed statistics, such as low precision due to the failure to differentiate cases with negation and differential diagnosis.

Usability. Besides understanding how NLP performs on multi-site EHR data, our ultimate goal is to learn from these errors and refine or re-train the model to improve its out-of-the-box performance. The error taxonomy can serve as a reference guide to enhance usability for NLP refinement (e.g., data pre-processing, model optimization, regex, and rulesets design). Based on the multi-site evaluation results, we re-grouped individual error classes into six higher-level error dimensions. As Figure 6 suggests, each study presents a unique error distribution. We can effectively identify the primary error types and design potential error mitigation strategies for each study. For example, all studies exhibited issues with the quality of corpus annotation, which can be addressed through quality control methods such as improved annotator training, double annotation, and iterative measurement of inter-annotator agreements. In the breast cancer study, the error analysis indicated a handful of linguistic errors, which may be resolved by improving

training data representation through different data sampling or increasing data size. However, systematically establishing the validity and reliability of the linkages between error classes and mitigation strategies requires additional evaluation studies. We hope to partner with more researchers and the informatics community to extend this functionality.

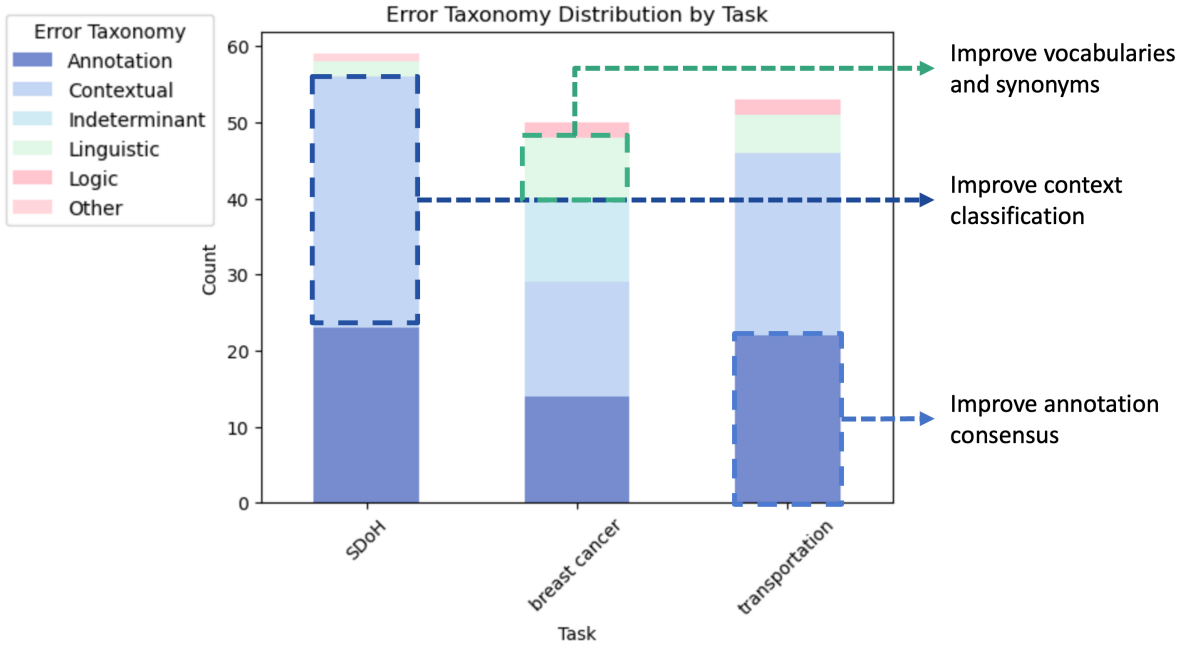

Figure 6. Visual Comparison of Error Class Across Three Differences

Challenges and limitations. We encountered several challenges in our study. First, the nature and prerequisites of error analysis require evaluators to be familiar with both the clinical task and NLP, posing additional challenges in identifying appropriate evaluators. In some specialized tasks, such as interpreting the difference between white matter disease and silent cerebral infarction, the task can only be performed by highly specialized neurologists or neuroradiologists who may lack knowledge about NLP models. Therefore, in some scenarios, error analysis may require collaboration between two domain experts. Additionally, this process often demands significant effort for initial task onboarding training and consensus development. Not only is the nature of the task more challenging compared to traditional corpus annotation tasks, but not all errors can be fully interpreted, and some errors may have multiple interpretations. For example, an NLP model failed to extract the expression 'family stated patient was ok at 10:00 AM, but around 3:00 PM the patient was confused and not making sense' may be due to both 'Subject' and 'Negation.' Despite

the challenging nature of the task, incorporating best practice guides and informatics tooling may help improve the quality and usability.

Having standardized annotation guidelines with detailed definitions, examples, and error analysis instructions can greatly improve standardization and consistency. In our prior study, we developed the TRUST process[26, 27] to enhance the reproducibility and validity of the clinical corpus annotation process, followed by the Findable, Accessible, Interoperable, and Reusable (FAIR) and Reproducible, Implementable, Transparent, and Explainable (RITE) principles[32]. A similar process should be considered when performing error analysis. Important components to be considered in the process include having an error analysis guideline (see Appendix 5 for example), organizing training sessions, using standardized evaluation tools, and engaging in team science collaboration. Proper training and education can help reduce process inconsistency and increase transparency, especially for a cross disciplinary team.

However, the current process for conducting systematic error analysis is still time-consuming and labor-intensive, even with the support of standardized guidelines. From a tooling perspective, future researchers could explore the potential direction of developing automated or semi-automated methods to assist in the classification and standardization of error analysis. One immediate opportunity with the rapid development of Large Language Models (LLMs) and domain-specific models may be helpful to assist with semi-automated error analysis. LLMs have demonstrated promising results in contextual understanding of clinical text. By combining a human-in-the-loop process[34] and Retrieval-Augmented Generation (RAG)[35] and optimizing the output using a knowledge base (e.g., error taxonomy and guidelines), the efficiency and effectiveness of the traditional error analysis workflow can be improved.

Future work. Our initial effort focuses on the comprehensiveness of the taxonomy. We plan to partner with communities such as ENACT and OHDSI NLP working groups and leverage a co-design workshop study[36] to assess and improve the usability, reliability, and applicability of the taxonomy.

Authors Contributions

S.F., J.S., H.L., and S.S. conceptualized and designed this study. S.F. and S.S. wrote a draft of the manuscript and conducted the analysis. H.H. performed the visualization analysis. S.F., L.W., and A.W. participated taxonomy consensus development and revision. N.Z. implemented the OWL representation. A.K., F.L., S.Z., R.Z., C.L., and Y.W. conducted the multisite evaluation. All authors critically revised the paper and provided important intellectual content. All authors approved this study.

Funding

This work was supported by National Institute on Aging R21 AG58738, National Institute on Aging R01 AG34676, and National Institute on Aging R01 AG068007.

Conflicts of Interest

All authors have no competing interests to declare.

Data Availability

Detailed definitions for each error class and properties can be accessed at https://github.com/OHNLP/ErrorAnalysis. The MedTator text annotation tool can be accessed at https://ohnlp.github.io/MedTator/.

Reference

1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. Journal of biomedical informatics. 2018;77:34-49.
2. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. Journal of Biomedical Informatics. 2020:103526.
3. Cusick M, Velupillai S, Downs J, Campion Jr TR, Sholle ET, Dutta R, Pathak J. Portability of natural language processing methods to detect suicidality from clinical text in US and UK electronic health records. Journal of affective disorders reports. 2022;10:100430.
4. Adekkanattu P, Jiang G, Luo Y, Kingsbury PR, Xu Z, Rasmussen LV, et al., editors. Evaluating the portability of an NLP system for processing echocardiograms: a retrospective, multi-site observational study. AMIA Annual Symposium Proceedings; 2019: American Medical Informatics Association.
5. Han P, Fu S, Kolis J, Hughes R, Hallstrom BR, Carvour M, et al. Multicenter Validation of Natural Language Processing Algorithms for the Detection of Common Data Elements in Operative Notes for Total Hip Arthroplasty: Algorithm Development and Validation. JMIR medical informatics. 2022;10(8):e38155.
6. Wu T, Ribeiro MT, Heer J, Weld DS, editors. Errudite: Scalable, reproducible, and testable error analysis. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019.
7. Yuan J, Vig J, Rajani N, editors. Isea: An interactive pipeline for semantic error analysis of nlp models. 27th International Conference on Intelligent User Interfaces; 2022.
8. Tang B, Cao H, Wu Y, Jiang M, Xu H, editors. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC medical informatics and decision making; 2013: BioMed Central.
9. Esuli A, Marcheggiani D, Sebastiani F. An enhanced CRFs-based system for information extraction from radiology reports. Journal of biomedical informatics. 2013;46(3):425-35.
10. Hoogendoorn M, Szolovits P, Moons LM, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artificial intelligence in medicine. 2016;69:53-61.
11. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, Ohe K. Extraction of adverse drug effects from clinical records. MEDINFO 2010: IOS Press; 2010. p. 739-43.
12. Doan S, Xu H, editors. Recognizing medication related entities in hospital discharge summaries using support vector machine. Proceedings of COLING International conference on computational linguistics; 2010: NIH Public Access.
13. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. Journal of the American Medical Informatics Association. 2010;17(4):383-8.
14. Yoon H-J, Roberts L, Tourassi G, editors. Automated histologic grading from free-text pathology reports using graph-of-words features and machine learning. 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); 2017: IEEE.
15. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics. 2015;53:196-207.
16. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics. 2017;75:S34-S42.
17. Wu Y, Xu J, Jiang M, Zhang Y, Xu H, editors. A study of neural word embeddings for named entity recognition in clinical text. AMIA annual symposium proceedings; 2015: American Medical Informatics Association.
18. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PloS one. 2018;13(2):e0192360.
19. Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. Journal of the American Medical Informatics Association. 2009;16(4):571-5.
20. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. The Journal of bone and joint surgery American volume. 2019;101(21):1931.

21. Nelson RE, Grosse SD, Waitzman NJ, Lin J, DuVall SL, Patterson O, et al. Using multiple sources of data for surveillance of postoperative venous thromboembolism among surgical patients treated in Department of Veterans Affairs hospitals, 2005–2010. Thrombosis research. 2015;135(4):636-42.
22. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, Chapman WW. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. Journal of Biomedical Semantics. 2016;7:1-12.
23. Fu S, Lopes GS, Pagali SR, Thorsteinsdottir B, LeBrasseur NK, Wen A, et al. Ascertainment of delirium status using natural language processing from electronic health records. The Journals of Gerontology: Series A. 2022;77(3):524-30.
24. Fu S, Thorsteinsdottir B, Zhang X, Lopes GS, Pagali SR, LeBrasseur NK, et al. A hybrid model to identify fall occurrence from electronic health records. International journal of medical informatics. 2022;162:104736.
25. Fu S, Lopes GS, Pagali SR, Thorsteinsdottir B, LeBrasseur NK, Wen A, et al. Ascertainment of Delirium Status Using Natural Language Processing From Electronic Health Records. The Journals of Gerontology: Series A. 2020;77(3):524-30.
26. Fu S. TRUST: Clinical Text Retrieval and Use towards Scientific Rigor and Transparent Process: University of Minnesota; 2021.
27. Fu S, Leung LY, Raulli A-O, Kallmes DF, Kinsman KA, Nelson KB, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC medical informatics and decision making. 2020;20(1):1-12.
28. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. Journal of the American Medical Informatics Association. 2022;29(7):1208-16.
29. Zhou S, Wang N, Wang L, Sun J, Blaes A, Liu H, Zhang R. A cross-institutional evaluation on breast cancer phenotyping NLP algorithms on electronic health records. Computational and Structural Biotechnology Journal. 2023;22:32-40.
30. Liu S, Wen A, Wang L, He H, Fu S, Miller R, et al. An open natural language processing (NLP) framework for EHR-based clinical research: a case demonstration using the National COVID Cohort Collaborative (N3C). Journal of the American Medical Informatics Association. 2023:ocad134.
31. Fu S, Vassilaki M, Ibrahim OA, Petersen RC, Pagali S, St Sauver J, et al. Quality assessment of functional status documentation in EHRs across different healthcare institutions. Frontiers in Digital Health. 2022;4.
32. Fu S, Wang L, Moon S, Zong N, He H, Pejaver V, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review. Clinical and translational science. 2023;16(3):398-411.
33. Friedman CP, Wyatt J. Evaluation methods in biomedical informatics: Springer Science & Business Media; 2005.
34. Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Generation Computer Systems. 2022;135:364-81.
35. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.
36. Wang Q, Jing S, Goel AK, editors. Co-designing AI agents to support social connectedness among online learners: functionalities, social characteristics, and ethical challenges. Proceedings of the 2022 ACM Designing Interactive Systems Conference; 2022.

Figure Legends

Figure 1. Overview of Error Taxonomy Development, Evaluation, and Dissemination Process. Site 1: a network of community-based health systems, Site 2: a nonprofit community-based hospital, Pitt: The University of Pittsburgh, Umass: University of Massachusetts Chan Medical School, UMN: The University of Minnesota.

Figure 2. Overview of Error Taxonomy in OWL Format.

Figure 3. Comparison of Error Type Distribution between False Positive (Left) and False Negative (Right) Cases in Large Corpus Evaluation.

Figure 4. Visual Comparison of Error Class Across Three Differences.

Figure 5. Visualization of Error Analytics on based on MedTator.

Figure 6. Visual Comparison of Error Class Across Three Differences.