Extraction of Crohn's Disease Clinical Phenotypes from Clinical Text Using Natural Language Processing

Linea Schmidt[a,b,c,*], Susanne Ibing[a,b,c,1,*], Florian Borchert[a], Julian Hugo[a,b,c], Allison Marshall[d], Jellyana Peraza[e], Judy H. Cho[e], Erwin P. Böttinger[a,b,c], Ryan C. Ungaro[f,1]

[a] Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Germany
[b] Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, USA
[c] Windreich Dept. of Artificial Intelligence & Human Health, Icahn School of Medicine at Mount Sinai, New York, USA
[d] Department of Medicine, Mount Sinai Health System, New York, USA
[e] Department of Pathology, Molecular, and Cell Based Medicine, Icahn School of Medicine at Mount Sinai, New York, USA
[f] The Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, USA

[1] Corresponding authors:
Susanne Ibing, susanne.ibing@hpi.de
Ryan C. Ungaro, ryan.ungaro@mssm.edu

[*] These authors contributed equally to this work.

Abstract

Background: Crohn's Disease (CD) patient heterogeneity in clinical practice is captured by the Montreal Classification. While the underlying concepts, disease behavior and age at diagnosis, are relevant outcomes and covariates in studies from real-world data, extracting this clinical information through manual chart review is labor-intensive and with limited scalability.

Methods: We developed and evaluated automated phenotyping algorithms to extract disease behavior and age at diagnosis from clinical narrative texts, using a rule-based approach based on the spaCy framework, and an approach based on zero-shot inference. The underlying data included 49,572 clinical notes and 2,204 radiology reports from 584 CD patients of the Mount Sinai Crohn's and Colitis Registry. A test set of 200 clinical texts per classification category was labeled at sentence-level, in addition to patient-level ground truth data. The algorithms were evaluated based on their recall, precision, specificity values, and F1-scores.

Results: For the labeled dataset, an overall Cohen's kappa inter-annotator agreement of 0.84 was achieved. The rule-based approach yielded high recall and precision values (0.75 - 1.00) on a note level for the behavioral disease phenotype using clinical notes, with slightly reduced performance using radiology reports. For age at diagnosis, recall and precision values of 0.81 and 0.88 were achieved on note-level, respectively. For both categories, the performance on patient- compared to note-level was reduced, potentially due to the accumulation of false positives and limitations in the data availability.

Conclusion: Based on our newly annotated dataset, we demonstrated the feasibility of automatically extracting disease behavior and age at diagnosis from clinical text. The resulting labels may facilitate extensive cohort analyses based on electronic health records, and support chart review processes in the future.

Key words: Crohn's Disease, Natural Language Processing, Automated Phenotyping, Montreal Classification

1 Introduction

Crohn's Disease (CD), one of the main entities of Inflammatory Bowel Disease (IBD), is an immune-mediated disease marked by recurrent episodes of chronic inflammation of the Gastrointestinal (GI) tract. The disease is characterized by high heterogeneity regarding disease course and treatment response [1]. Its progressive nature leads to the accumulation of disease complications and eventually surgical interventions, and thereby the build-up of structural alterations and irreversible bowel damage [2].

The Montreal Classification is used to group CD patients considering three categories: *age at diagnosis*, *disease location*, and *disease behavior* (Table 1) [3], [4]. Age at diagnosis refers to the age at initial CD diagnosis and disease location to the coarse region of inflammation. Disease behavior comprises different disease complications of CD, such as strictures (B2), fistulas, and abscesses (B3). Strictures are luminal narrowings in any part of the GI tract. They are developed due to chronic inflammation of the mucosa, resulting in excessive repairs of the area of inflammation and, eventually, the mixture of inflamed and scarred tissue [5]. Fistulas refer to abnormal passageways that form between different parts of the GI tract, between the GI tract and other organs, or between the GI tract and the exterior. They can develop due to chronic inflammation and damage to the intestinal wall [1], [6]. In the context of CD, an abscess is a localized accumulation of pus that can develop due to inflammation or infection, due to complicated and active disease, or after surgical interventions [7]. Perianal disease is regarded as a disease modifier that can co-occur with any of the other disease phenotypes (B1-B3): Any of the penetrating or stricturing disease complications that occur in the perianal region are counted as perianal disease. Disease behavior is not a static classification category, since disease complications can be gained during the course of the disease [2].

In a recent publication, an expert consensus of the European Crohn's and Colitis Organization (ECCO) discussed core outcomes and outcome measures that are relevant to be reported in IBD studies based on real-world data, such as Electronic Health Records (EHR) [8]. Even though randomized studies are considered gold-standard, studies based on real-world data are of interest to derive complementary evidence. The data is essential as it enables longitudinal analysis of rich clinical data beyond the natural limitation of clinical trials [9]. For instance, in CD, through regular health care interactions for disease monitoring of the chronic condition and various types of interactions, such as endoscopic procedures, lab tests, radiology imaging, and other regular encounters, rich clinical information can be derived from patients' EHR [10]. Furthermore, studies based on real-world data can mitigate the fact that many patients with severe disease courses are not included in clinical trials [11]. According to the ECCO expert group, when reporting on disease complications,

Table 1: Montreal Classification for Crohn's Disease patients according to Silverberg, Satsangi, Ahmad, *et al.* [3]. *L4 is a modifier that can be added to L1–L3 when concomitant upper gastrointestinal disease is present. †"p" is added to B1–B3 when concomitant perianal disease is present.

Category	**Classification**	**Definition**
Age at Diagnosis	A1	Below 16 years
	A2	Between 17 and 40 years
	A3	Above 40 years
Disease Location	L1	Ileal
	L2	Colonic
	L3	Ileocolonic
	L4*	Isolated upper disease
Disease Behavior	B1	Non-stricturing and non-penetrating
	B2	Stricturing
	B3	Penetrating
	p†	Perianal disease modifier

it is recommended to consider the presence of strictures, fistulas, and the disease phenotype as core outcomes. The Montreal Classification was recommended as outcome measure [8]. Next to disease behavior, age at diagnosis is part of the Montreal Classification and an important clinical component for CD clinical care and study cohort characterization, as it allows the deduction of disease duration, a prognostic factor for treatment response with biologics [12].

For large EHR-based studies, extracting clinical phenotypes in CD through chart review is a time-consuming process, requiring a lot of manual labor of domain experts. Structured EHR are not reliable enough to extract the relevant information: Ananthakrishnan *et al.* showed that in the health records of 399 CD patients, based on clinical notes, 36% of the cohort were affected by fistulizing disease while only being coded in 12%. Furthermore, 40% of the cohort had stricturing disease according to information extracted from clinical text, but it was only coded for 25%. Only perianal disease was coded in a higher fraction of patients (13%) than being mentioned in the narrative text (11%) [13].

The nuanced and often complex descriptions of the behavioral phenotype in clinical text include descriptions of symptoms and treatment responses, as well as the progression of the disease. Since this data is stored mainly in clinical narrative text, a patient's disease behav-

ior is usually extracted by manual chart review [14]–[16]. Automated phenotyping based on Natural Language Processing (NLP) techniques, including information from clinical notes, could facilitate patient classification on a large scale with minimal manual labeling required. For instance, Stidham *et al.* demonstrated the successful extraction of extraintestinal manifestations in IBD patients recently using a rule-based NLP approach [17].

Clinical phenotyping algorithms can typically be divided into two main categories: rule-based and Machine Learning (ML)-based approaches. The choice between these approaches relies on factors like the availability of data and the complexity of the task at hand [18]. Rule-based techniques utilize predefined rules or criteria, often defined by experts and based on diagnostic codes, clinical test results, medications, or other clinical data. Creating these rules typically necessitates expertise in the relevant field. On the contrary, ML-based methods employ ML algorithms to recognize patterns in the data corresponding to different phenotypes. These algorithms can handle vast amounts of data and identify intricate patterns that might be challenging for humans to discern. Only a few years ago, most studies have been conducted using concept extraction. However, more deployment of unsupervised techniques is on the rise [19]. In their benchmark paper, Moldwin *et al.* demonstrated, amongst others, for digestive diseases, that the incorporation of unstructured data outperforms models that are only based on structured EHR [20]. Furthermore, for the identification of lumbar spine imaging findings, a developed ML system outperformed the rule-based NLP approach [21]. Nevertheless, a rule-based approach provides increased transparency compared to ML approaches, particularly when applying Large Laguage Models (LLMs).

In this work, we describe the development of a novel, sentence-based labeled dataset, including annotations of disease phenotype and age at diagnosis in clinical notes of CD patients. We used this dataset to develop and evaluate rule-based phenotyping algorithms and compare them with a baseline zero-shot transformer model in case of the behavioral disease phenotype. The established pipeline facilitates the large-scale labeling of previously unknown clinical narrative text.

2 Materials and Methods

2.1 Data Collection and Preprocessing

Clinical notes from the EHR of the Mount Sinai Data Warehouse (MSDW) [22] were acquired via the Artificial Intelligence Ready Mount Sinai (AIR·MS) platform. This dataset was further enriched with reports from the radiology department, allowing the inclusion of Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) reports. A total of 792 CD patients from the Mount Sinai Crohn's and Colitis Registry (MSCCR) were consid-

ered for this study [14]. We preprocessed the available clinical notes and radiology reports by removing irrelevant note types (e.g., telephone encounter and patient instructions) and texts that did not contain a CD-related context (Figure 1). 584 of the 792 patients had at least one non-empty clinical note available after filtering. The clinical text dates range from February 1940 (first clinical text) to May 2023 (latest clinical text). For the annotation and extraction of the age at diagnosis, we additionally filtered notes for regular expressions containing key expressions such as "diagnosed", "Crohn's [...] since", or "age at"[1]. To allow for further granularity, we split all clinical texts into sentences (Figure 1).

2.2 Annotation and Dataset Creation

For disease behavior, annotation guidelines were based on the COMPASS-IBD study and the Ocean State Crohn's and Colitis Area Registry (OSCCAR) data dictionary [15], [16]. Two annotators, an internal medicine resident and an IBD researcher, labeled 200 notes on a sentence level (Figure 2). An agreement sample of 50 notes (5,543 sentences) was used to ascertain the Inter-Annotator Agreement (IAA), which was measured using Cohen's kappa statistics [23]. After resolving disagreements, a curated dataset was created and used as a test set. Additionally, a development set consisting of 200 clinical texts was labeled by non-experts. Rules were exclusively developed using this development dataset and evaluated on previously unseen test data. To allow an evaluation of the disease behavior on the patient level, we used a labeled subset from MSCCR. The data comprised 134 labeled patients with available clinical texts until their first endoscopy within the MSCCR study.

The annotation of the age at diagnosis was conducted by the IBD researcher, labeling 200 additional clinical texts on a sentence level based on three categories: age at diagnosis, diagnosis year, and disease duration. Notably, the current ground-truth age at diagnosis was calculated using patients' birth years and the note dates. After curation, this dataset was split into a test and a development set. For a patient-level evaluation, the original labels from MSCCR patients with available clinical texts were used.

[1]Included clinical texts needed to match at least one of the following patterns:
`(D|d)iagnosed|DIAGNOSED`,
`((C|c)rohn|CROHN|cd|CD)[^a-zA-Z0-9]*(since|SINCE)`,
`(D|d)isease[^a-zA-Z0-9]*(O|o)nset|DISEASE[^a-zA-Z0-9]*ONSET`,
`(A|a)ge[^a-zA-Z0-9]*(A|a)t[^a-zA-Z0-9]*(D|d)iagnosis`,
`AGE[^a-zA-Z0-9]*AT[^a-zA-Z0-9]*DIAGNOSIS`

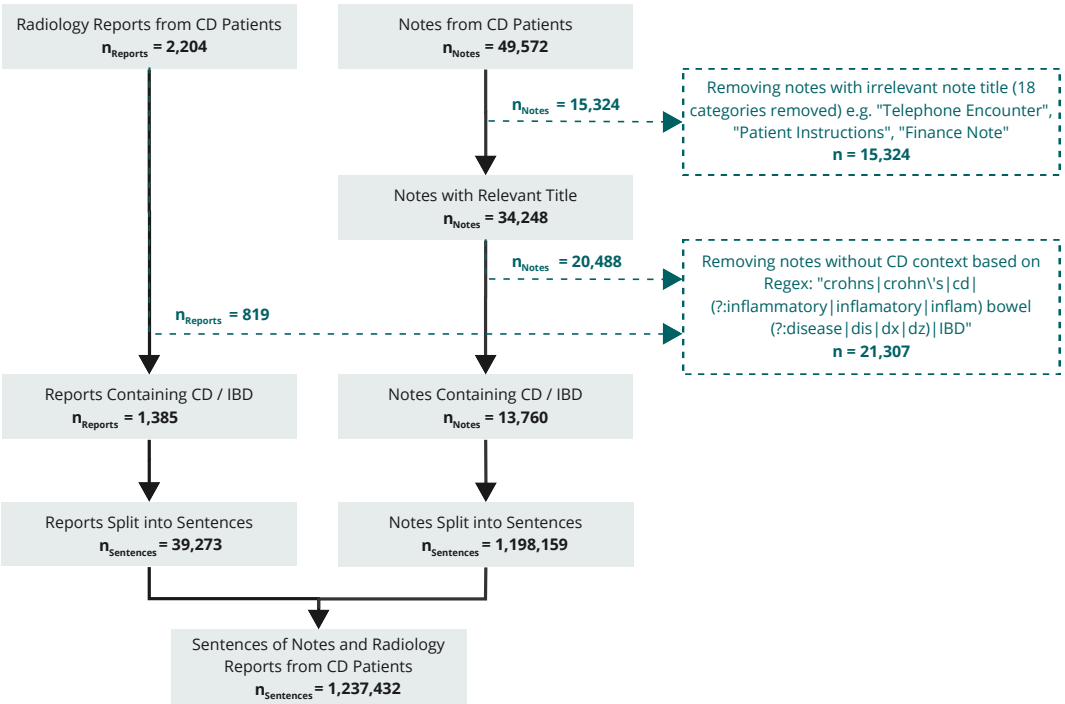

Figure 1: Data sources and preprocessing steps. After extracting all available clinical notes from CD patients in MSCCR up until two weeks after the date of initial endoscopy and biopsy for sample collection for the study, all notes with irrelevant tiles were removed. Subsequently, from the available clinical notes and radiology reports, only disease-relevant texts were further processed by splitting them into sentences.

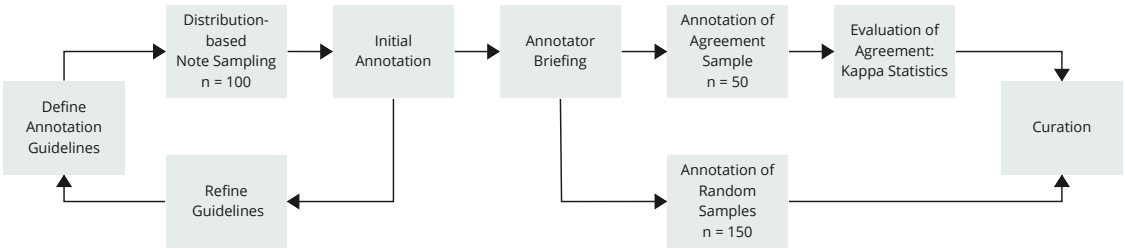

Figure 2: Labeling process. The process from building annotation guidelines to a final annotated and curated dataset for CD disease complications containing 150 clinical notes and 50 radiology reports.

2.3 Disease Behavior Phenotyping

Two primary methods for disease behavior phenotyping were adopted: a zero-shot classifier approach and a rule-based algorithm.

2.3.1 Zero-Shot Classifier

The disease behavior classification task on the sentence level was framed as a Natural Language Inference (NLI) problem, enabling the usage of the pre-trained model *facebook/bart-large-mnli* model[2] from the Hugging Face Hub [24]. Multiple classifiers targeting specific behavioral phenotypes were utilized to identify instances of B2 (used classification labels: "structuring", "narrowing") and B3 (used classification labels: "abscess", "fistula") and the perianal region (used classification labels: "perianal", "rectal", "ileorectal"). A score threshold of >0.7 was set for categorizing conditions. In case B2 or B3 matched a sentence and the perianal region, the match was treated as a perianal disease match. An additional classifier for uncertainty detection was used by passing "uncertainty" as a label to the pre-trained Large Language Model (LLM). Sentences exceeding an "uncertainty" score of 0.7 were discarded.

2.3.2 Rule-based Approach

Grounded on the Clinical Informatin Extraction for Phenotyping and Predictive modeling using EHR (CLIPPEHR) infrastructure, the rule-based approach leverages spaCy [25], scispaCy [26], and medspaCy [27]. A custom spaCy component, BehavioralPhenoCategorizer, is constructed for phenotype extraction. The *en_core_sci_md* scispaCy model is the cornerstone for syntactic analyses and named entity recognition. After preprocessing, abbreviation detection, and Unified Medical Language System (UMLS) linking using a curated subset of UMLS Metathesaurus codes, patterns were established to detect specific CD behavioral phenotypes. The development process utilized spaCy's Matcher class to design patterns that describe token sequences for accurate disease phenotyping of CD. Multiple patterns were crafted: for specific phenotype complications, direct string-level matches, UMLS linkages, and two additional patterns addressing medical conjectures (uncertainty matcher) and explicit exclusions (exclusion matcher). These patterns were further refined to differentiate complications like B2/B3 from perianal disease through UMLS linking and token-level regular expression-like patterns.

In clinical texts, often both the presence and absence of medical conditions are described, making effective negation detection crucial. For behavioral phenotyping in CD, two

[2]https://huggingface.co/facebook/bart-large-mnli

strategies were adopted: one leveraging medspaCy — a rule-based approach that identifies negation patterns and uses dependency parsing to determine negated entities, and the other utilizing a Transformer-based Clinical Assertion and Negation Classification BERT model [28]. For the latter, we deployed the pre-trained *bvanaken/clinical-assertion-negation-bert* model from the Hugging Face Hub[3] that detects entity absence with a probability score, considering spans as negated if they surpass a 0.5 threshold.

The BehavioralPhenoCategorizer processes each document in stages: initial categorization using UMLS matching, pattern application to detect matches, followed by exclusion checks based on direct string matching of terms such as "no" or "not" and the results of the chosen negation detection method. In case of a CD complication match, a context window of up to seven tokens is scanned for uncertainty or exclusion patterns. If the match is not determined to be negated but still linked to B2 or B3 classifications, proximity to mentioning the perianal region is checked, leading to potential phenotype reassignment. Phenotypes are stored and aggregated at different levels (patient, note, or sentence), with the most severe phenotype following the order B1<B2<B3. If the input data contains labeled ground-truth information, the phenotyping performance is assessed (Figure 3).

2.4 Age at Diagnosis Phenotyping

Similar to the disease behavior, a custom spaCy component, AgeAtDiagnosisClassifier, was engineered to determine the age at diagnosis. Through a series of pattern matching, textual spans indicating age at diagnosis (e.g., "diagnosed with 8 years"), disease duration (e.g., "CD since 10 years"), or diagnosis year (e.g., "CD diagnosed in 2002") were recognized. Subsequent analysis determined the age at diagnosis based on these matches, the patient's year of birth, and the date of the note. Of note, the identified year of birth includes an error margin of ± 1 year, since exact dates are not frequently mentioned in clinical notes. For performance metrics, True Positives (TP) corresponded to accurately identified ages at diagnosis (within $\pm$1-year). True Negatives (TN) were accurate identifications where age information was absent, False Positives (FP) were defined as wrongly identified age at diagnosis labels, and False Negatives (FN) represented overlooked labeled instances.

2.5 Evaluation of the Algorithms Performance

We evaluated the algorithms based on recall, precision, specificity and F1-Score. During the development of the disease behavior rule-based phenotyping algorithm, our primary metric of interest was recall, given the importance of the sensitive identification of positive

[3]https://huggingface.co/bvanaken/clinical-assertion-negation-bert

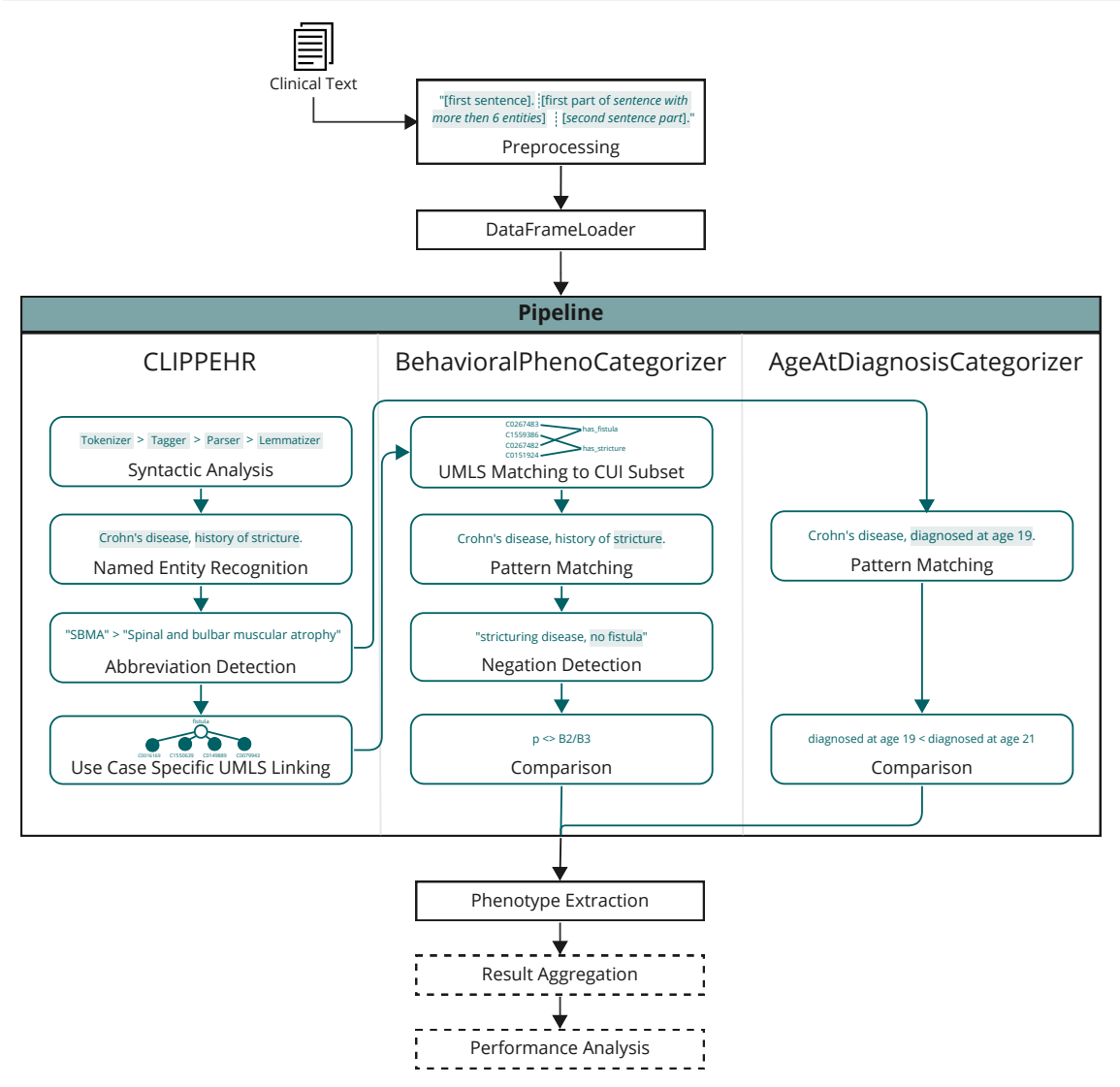

Figure 3: Rule-based phenotyping algorithm. The phenotyping process starts with clinical texts from radiology and clinical notes as input. The spaCy pipeline contains, on the one hand, elements defined by CLIPPEHR and, on the other hand, the newly developed BehavioralPhenoCategorizer or the AgeAtDiagnosisCategorizer. After the phenotype extraction, result aggregation and performance analysis are optional. Evaluation is only conducted if ground-truth labels are available.

instances. For age at diagnosis, on the other hand, we prioritized precision to focus on the accurate extraction of the information for downstream tasks. Maintaining a balanced precision, F1-Score, and specificity were set as secondary aims.

3 Results

We developed and evaluated phenotyping algorithms for two dimensions of the Montreal Classification, *disease behavior* and *age at diagnosis*, that are being used to group CD patients according to their clinical presentation and disease history. Two raters were included in the annotation process for a more complex and nuanced description of the disease behavior category to evaluate IAA. In the following, we separately present the results for disease behavior and age at diagnosis extraction.

3.1 Automated Extraction of the Disease Behavior

To evaluate the performance of the disease behavior phenotyping algorithms, we created a newly annotated dataset comprising 150 clinical notes and 50 radiology reports, with a total of 15,390 sentences (Table 2).

50 of these clinical notes were handed to two different annotators. The quality of the labeling process was determined through Cohen's kappa agreement scores. We observed an overall IAA score of 0.84 (B1: 0.83; B2: 0.81; B3: 0.85; perianal disease: 0.87). These results indicate a near-perfect consensus among annotators [29], underlining the robustness and appropriateness of the labeled data as ground truth for subsequent phenotyping algorithm evaluations.

After the annotators found a consensus for all disagreement instances, in the finalized, curated test set, approximately 1% of clinical note sentences and 3.6% of radiology report sentences got a B2 or B3 label assigned. This aligns with the perianal disease modifier, with roughly 0.8% of clinical note sentences and 2.1% of radiology report sentences with a positive annotation (Table 2).

With the zero-shot classifier based on the pre-trained NLI model as baseline analysis, we were challenged to balance the various performance metrics. Notably, while a satisfactory recall score of approximately 0.81 for B3 was attained based on clinical notes, there was a consequent compromise in precision and F1-Score, with values ranging from below 0.01 to 0.46 on sentence-level (Suppl. **??**). We observed a congruent pattern when evaluating radiology reports (Suppl. **??**). Of note, the model was not fine-tuned with our specific classification tasks. Since a high number of false positives was already predicted on sentence level, we did not further aggregate and evaluate the results on note-level.

Table 2: Overview of the annotation process of the test dataset for the behavioral disease phenotype using clinical notes and radiology reports (rep.).

		Clinical Notes	Radiology Rep.	Total
Total number of notes		150	50	200
Total number of sentences		14,236	1,154	15,390
Mean sentences per note (SD)		95 (±87)	23 (±11)	-
Not B2/B3	**Notes**	112	32	144
	Sentences	14,094	1,113	15,207
B2	**Notes**	13	7	20
	Sentences	62	24	86
B3	**Notes**	25	11	36
	Sentences	80	17	97
Perianal disease	**Notes**	25	7	32
	Sentences	113	24	137

The rule-based approach was evaluated on patient-, note-, and sentence-level. First, to optimize our pipeline, we conducted a series of experiments with varied settings regarding rule types and negation detection options on the test set at the sentence level. Concerning the differentiation of rule types, we observed that a synergistic approach combining UMLS matching rules with rules for direct string matching was superior in its performance compared to applying either of the rule types alone. Notably, the exclusive employment of string matching exhibited superior results compared to solely relying on UMLS matching across both clinical notes and radiology reports (Suppl. **??**). For negation detection, we analyzed the number of false positives and false negative disease complications using either the medspaCy negation detection component or the LLM Negation Classifier or no negation detection at all. The LLM Negation Classifier performed as the superior compared to the other two options, manifesting the lowest incidence of false negatives while preserving a substantial number of accurately identified instances (Suppl. **??**, Suppl. **??**).

With the automated behavioral phenotyping based on clinical notes, we yielded high recall values on note-level, ranging from 0.92 - 1.00 depending on the phenotype (Table 3). In particular, for perianal disease and B3, all instances were correctly identified. Based on radiology reports, these values dropped to 0.64 - 1.00, with less sensitive identification, in particular of B3 and B2. The underlying reasons may be the incorrect identification of B2 or B3 as perianal disease. Overall, the precision values and F1-Scores indicate over-classification, resulting in false positive disease complication labels.

For 134 patients of the MSCCR study, we extracted the disease phenotype at study

Table 3: Performance of different rule-based phenotyping algorithms on note level using the newly annotated test dataset.

Model	Phenotype	Recall	Precision	F1-Score	Specificity
Disease Behavior Model on Clinical Notes	**Not B2/B3**	0.94	1.00	0.97	1.00
	B2	0.92	0.75	0.83	0.97
	B3	1.00	0.86	0.93	0.97
	p - Yes	1.00	0.86	0.93	0.86
	p - No	0.97	1.00	0.98	1.00
Disease Behavior Model on Radiology Reports	**Not B2/B3**	0.91	0.94	0.92	0.89
	B2	0.71	0.50	0.59	0.88
	B3	0.64	0.78	0.70	0.95
	p - Yes	1.00	0.80	0.89	0.80
	p - No	0.95	1.00	0.98	1.00
Age at Diagnosis Model on Clinical Notes & Radiology Reports	**Age at Diagnosis**	0.81	0.88	0.85	0.68

enrollment through manual chart review, considering all clinical information up until this time point. Compared to the note-level analysis, we achieved a recall value of 0.71 and a precision of 0.48 for detecting any complication (B2 or B3). For the detection of perianal disease, the recall was 0.85 and precision 0.56, considerably decreased compared to the note-level analysis (Table 4). Of note, for the annotation process of the patient-level ground-truth labels, as the primary clinical information system was used, the basis of underlying data differed from the information available for the automated phenotyping.

3.2 Automated Extraction of the Age at Diagnosis

We evaluated the performance of the age at diagnosis extraction through the rule-based model on patient- and note-level. While on note-level we observed balanced performance metrics, exemplified by a recall of 0.81 and a precision of 0.88 (Table 3), on patient-level, a similar balance of the performance measures was achieved with slightly lower overall performance values (Table 4). The underlying data of the 584 MSCCR patients for evaluation on

patient-level was not specifically scanned for availability of the information in scope within the written clinical text. Therefore, the reduced performance of the model on patient-level may be explained by the limited data availability.

Table 4: Performance of different rule-based phenotyping algorithms on patient-level. 134 patients of the MSCCR cohort had available information on the behavioral disease phenotype through manual chart review. The age at diagnosis was evaluated on 584 MSCCR patients with available clinical narrative texts using the available age at diagnosis labels of the cohort.

Model	Phenotype	Recall	Precision	F1-Score	Specificity
Disease Behavior Model	**Not B2/B3**	0.65	0.83	0.73	0.48
	B2/B3	0.71	0.48	0.58	0.83
	p - Yes	0.85	0.56	0.68	0.56
	p - No	0.83	0.96	0.89	0.96
Age at Diagnosis Model	**Age at Diagnosis**	0.73	0.70	0.71	-

4 Discussion

In this work, we demonstrate the feasibility of the automatic extraction of disease behavior and age at diagnosis, two components of the Montreal Classification, from clinical texts using a rule-based NLP approach. To our knowledge, we are the first to describe phenotyping algorithms for the stated task.

We created two labeled datasets to evaluate our algorithm, including sentence-level annotations for disease behavior and age at diagnosis. In particular, the more complex descriptions of the disease phenotype required the evaluation of the two annotators based on the Cohen's kappa value. The IAA described in this work, except for perianal disease, surpass the kappa statistics documented in previous literature [30], [31]. Of note, the sentence-level annotation agreements are difficult to be compared with patient-level annotations. Overall, our results suggest the acceptability of our annotated data as a test dataset.

Shrestha *et al.* described in their work the identifications of disease phenotypes by using International Classification of Diseases (ICD) codes of the Swedish National Patient Register. Their reported recall values lie between 0.62 for B2/B3, 0.75 for B1, and 0.81 for

perianal disease, and on average 0.94 for the different phenotype groups of age at diagnosis [32]. While on a patient-level, their results were superior compared to ours, we have to acknowledge that with the fragmented health care systems as they exist in the US [33], coded information in EHR data is not reliable enough to extract the complex clinical information [13]. We developed a rule-based pipeline working directly with clinical text to mitigate this lack of coded information. On a note-level, we achieve recall values between 0.64 and 1.00, and precision values between 0.50 and 1.00. The overall performance is similar as described for other tasks in the literature, for instance, the extraction of extraintestinal manifestations [17].

Our models for disease behavior classification performed superior on clinical notes compared to radiology reports (Table 3). One underlying reason might be the fact that radiology reports were notably underrepresented in both our test and development sets. Characteristically, radiology reports tend to exhibit longer sentences, employ more intricate language structures, contain fewer spelling errors, and frequently include suggestions, exclusions, and negations. These different text compositions may have influenced the overall performance, highlighting the need for broader representation and potentially different processing strategies for diverse report types in future studies.

For disease behavior, we noticed a trend towards over-classification on the note- and patient-level. Through error analysis on a note-level, we realized that a major error source was the misclassification of B2/B3 and perianal disease. The description of the penetrating or stricturing disease complication can be the same in these cases since the localization of the disease complications defines the correct phenotype. Therefore, this classification task is of particular challenge.

For a more in-depth understanding of false classifications on a patient level, we analyzed five falsely positive and falsely negative classified patients for B2/B3 and perianal disease. False positive instances for B2/B3 mainly arose from the description of similar complications in other disease contexts (e.g., carotid stenosis), complex sentence structures leading to errors in negation detection, and confusion with perianal disease. Instances misclassified as perianal disease are suspected to be partly wrong-labeled, and in one instance the negation detection was not sophisticated enough to catch the negation in the given sentence structure. For the patients with false negative labels of B2/B3 and perianal disease, there was no clear description of the phenotype in the clinical texts.

For the age at diagnosis extraction, challenges mainly arose due to the varied representations of dates in the data, coupled with the task of unambiguously linking a date occurrence to the diagnosis of CD.

While our study shows promising results, we have to acknowledge certain limitations. Foremost, our findings highlight the difficulty of completely replacing manual chart review

with automated NLP-based phenotyping if the underlying data basis is not the same. In our case, due to the high fragmentation of clinical data into multiple IT systems, we did not have access to endoscopy reports for our study. Next to an increased development burden, this experienced limitation of the results to the available data is generally a typical drawback when it comes to rule-based algorithms [34].

Our models were evaluated on only Mount Sinai Health System (MSHS)-internal clinical texts. This poses a potential limitation, as the algorithms might be particularly tuned to language idiosyncrasies specific to physicians at Mount Sinai or the reporting conventions typical of this institution. With external evaluation, we may be able to make statements about the models' generalizability to other clinical settings with different linguistic nuances or documentation practices.

The presented rule-based models offer systematic and transparent reasoning and are thus potentially the preferred support for labeling tasks in a clinical setting, especially when no baseline for the stated problem exists. Nevertheless, based on the latest developments of ML-based approaches, they might offer potential for future improvements. On a technical level, the results of the zero-shot inference model should be considered preliminary. In particular, we used the off-the-shelf pre-trained solution to derive a baseline for comparison purposes, which was not a fine-tuned model on task-specific labeled data. Additionally, the NLI model was pre-trained on general-domain rather than clinical data. With the described solution, we cannot draw final conclusions on whether a rule-based approach can be generally preferred, but should further evaluate in-context or other few-shot learning approaches that are tailored to the biomedical domain like GatorTron [35]. Also, the capabilities of newer general models such as GPT-4 [36] would be of interest. Besides these potential limitations of the classifier, our method of sentence splitting occasionally resulted in phenotype information being fragmented across multiple sentences, complicating the classification task.

Nevertheless, with the described phenotyping algorithms and results, we are confident that we can contribute towards studies based on large cohorts and that our algorithms would be helpful to support chart review, accelerate the process of generating labels for, e.g., clinical studies, and would overall be beneficial to optimize label quality.

5 Conclusion

In this work, we successfully established a comprehensive phenotyping pipeline for two components of the Montreal Classification: disease behavior and age at diagnosis. We describe two newly developed custom spaCy components, the BehavioralPhenoCatego-

rizer and AgeAtDiagnosisCategorizer. The development and validation of our NLP-based pipeline were supported by creating a newly annotated dataset, which will serve as a resource for subsequent investigations. While our rule-based approaches have demonstrated high performance on note- and patient-level, further exploration into leveraging machine learning models, particularly LLM, is planned. We believe this would contribute to a more robust system capable of outperforming or complementing current methods. Our approach can serve as a strong baseline for such future developments. We anticipate that NLP-based information extraction for extensive cohort studies based on real-world data may increasingly be applied in the future.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

We are very greatful to all participants of the Mount Sinai Crohn's and Colitis Registry. Further, we thank Eugenia Alleva, and Jan Philipp Sachs for their support and fruitful discussions. We also thank Manbir Singh, Herve DiBello, and Lewis Lo for IT and data access support. Many thanks to Danielle Cohen, Alexandra Deutschenbaur, Narges Ghaedi Bardeh, Daniel Jühling, Lisa Koeritz, Larissa Röhrig, Marco Schaarschmidt, Jonathan Wilke, and Paul Wullenweber for their their development of the CLIPPEHR pipeline. This work is in part supported through the AI-Ready Mount Sinai (AIR.MS) research platform and the MSDW resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai. This work was funded by the Joachim-Herz Foundation (to SI), and the Hasso Plattner Foundation (HPF).

References

[1] J. Torres, S. Mehandru, J.-F. Colombel, and L. Peyrin-Biroulet, "Crohn's disease," en, *The Lancet*, vol. 389, no. 10080, pp. 1741–1755, Apr. 2017, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(16)31711-1.

[2] B. Pariente, J.-Y. Mary, S. Danese, *et al.*, "Development of the Lémann index to assess digestive tract damage in patients with Crohn's disease," eng, *Gastroenterology*, vol. 148, no. 1, 52–63.e3, Jan. 2015, ISSN: 1528-0012. DOI: 10.1053/j.gastro.2014.09.015.

[3] M. S. Silverberg, J. Satsangi, T. Ahmad, *et al.*, "Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology," eng, *Canadian Journal of Gastroenterology = Journal Canadien De Gastroenterologie*, vol. 19 Suppl A, 5A–36A, Sep. 2005, ISSN: 1916-7237. DOI: 10.1155/2005/269076.

[4] J. Satsangi, M. S. Silverberg, S. Vermeire, and J.-F. Colombel, "The Montreal classification of inflammatory bowel disease: Controversies, consensus, and implications," en, *Gut*, vol. 55, no. 6, pp. 749–753, Jun. 2006, Publisher: BMJ Publishing Group Section: Leading article, ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gut.2005.082909.

[5] F. Rieder, E. M. Zimmermann, F. H. Remzi, and W. J. Sandborn, "Crohn's disease complicated by strictures: A systematic review," en, *Gut*, vol. 62, no. 7, pp. 1072–1084, Jul. 2013, Publisher: BMJ Publishing Group Section: Recent advances in clinical practice, ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gutjnl-2012-304353.

[6] M. Scharl and G. Rogler, "Pathophysiology of fistula formation in Crohn's disease," *World Journal of Gastrointestinal Pathophysiology*, vol. 5, no. 3, pp. 205–212, Aug. 2014, ISSN: 2150-5330. DOI: 10.4291/wjgp.v5.i3.205.

[7] A. T. P. Carvalho, B. C. Esberard, and A. da Luz Moreira, "Current management of spontaneous intra-abdominal abscess in Crohn's disease," en, *Journal of Coloproctology*, vol. 38, no. 2, pp. 158–163, Apr. 2018, ISSN: 2237-9363. DOI: 10.1016/j.jcol.2016.05.003.

[8] J. Hanzel, P. Bossuyt, V. Pittet, *et al.*, "Development of a Core Outcome Set for Real-world Data in Inflammatory Bowel Disease: A European Crohn's and Colitis Organisation [ECCO] Position Paper," *Journal of Crohn's and Colitis*, jjac136, Oct. 2022, ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjac136.

[9] L. Blonde, K. Khunti, S. B. Harris, C. Meizinger, and N. S. Skolnik, "Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician," *Advances in Therapy*, vol. 35, no. 11, pp. 1763–1774, 2018, ISSN: 0741-238X. DOI: 10.1007/s12325-018-0805-y. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6223979/ (visited on 10/11/2023).

[10] B. Veauthier and J. R. Hornecker, "Crohn's Disease: Diagnosis and Management," eng, *American Family Physician*, vol. 98, no. 11, pp. 661–669, Dec. 2018, ISSN: 1532-0650.

[11] C. Ha, T. A. Ullman, C. A. Siegel, and A. Kornbluth, "Patients Enrolled in Randomized Controlled Trials Do Not Represent the Inflammatory Bowel Disease Patient Population," *Clinical Gastroenterology and Hepatology*, vol. 10, no. 9, pp. 1002–1007, Sep. 2012, ISSN: 1542-3565. DOI: 10.1016/j.cgh.2012.02.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1542356512002017 (visited on 10/11/2023).

[12] S. Ben-Horin, L. Novack, R. Mao, *et al.*, "Efficacy of Biologic Drugs in Short-Duration Versus Long-Duration Inflammatory Bowel Disease: A Systematic Review and an Individual-Patient Data Meta-Analysis of Randomized Controlled Trials," *Gastroenterology*, vol. 162, no. 2, pp. 482–494, 2022, Publisher: The Authors, ISSN: 15280012. DOI: 10.1053/j.gastro.2021.10.037. [Online]. Available: https://doi.org/10.1053/j.gastro.2021.10.037.

[13] A. N. Ananthakrishnan, T. Cai, G. Savova, *et al.*, "Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: A Novel Informatics Approach," *Inflammatory bowel diseases*, vol. 19, no. 7, pp. 1411–1420, Jun. 2013, ISSN: 1078-0998. DOI: 10.1097/MIB.0b013e31828133fd.

[14] R. Kosoy, S. Kim-Schulze, A. Rahman, *et al.*, "Deep Analysis of the Peripheral Immune System in IBD Reveals New Insight in Disease Subtyping and Response to Monotherapy or Combination Therapy," en, *Cellular and Molecular Gastroenterology and Hepatology*, vol. 12, no. 2, pp. 599–632, Jan. 2021, ISSN: 2352-345X. DOI: 10.1016/j.jcmgh.2021.03.012.

[15] S. L. Gold, L. G. Rabinowitz, L. Manning, *et al.*, "High Prevalence of Malnutrition and Micronutrient Deficiencies in Patients With Inflammatory Bowel Disease Early in Disease Course," *Inflammatory Bowel Diseases*, vol. 29, no. 3, pp. 423–429, May 2022, ISSN: 1078-0998. DOI: 10.1093/ibd/izac102.

[16] B. E. Sands, N. LeLeiko, S. A. Shah, R. Bright, and S. Grabert, "OSCCAR: Ocean State Crohn's and Colitis Area Registry," eng, *Medicine and Health, Rhode Island*, vol. 92, no. 3, pp. 82–85, 88, Mar. 2009, ISSN: 1086-5462.

[17] R. W. Stidham, D. Yu, X. Zhao, *et al.*, "Identifying the Presence, Activity, and Status of Extraintestinal Manifestations of Inflammatory Bowel Disease Using Natural Language Processing of Clinical Notes," *Inflammatory Bowel Diseases*, vol. 29, no. 4, pp. 503–510, Apr. 2023, ISSN: 1078-0998. DOI: 10.1093/ibd/izac109. [Online]. Available: https://doi.org/10.1093/ibd/izac109 (visited on 10/11/2023).

[18] B. Khosravi, P. Rouzrokh, and B. J. Erickson, "Getting More Out of Large Databases and EHRs with Natural Language Processing and Artificial Intelligence: The Future Is Here," en-US, *JBJS*, vol. 104, no. Suppl 3, p. 51, Oct. 2022, ISSN: 0021-9355. DOI: 10.2106/JBJS.22.00567.

[19] H. Alzoubi, R. Alzubi, N. Ramzan, D. West, T. Al-Hadhrami, and M. Alazab, "A Review of Automatic Phenotyping Approaches using Electronic Health Records," en, *Electronics*, vol. 8, no. 11, p. 1235, Nov. 2019, Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2079-9292. DOI: 10.3390/electronics8111235.

[20] A. Moldwin, D. Demner-Fushman, and T. R. Goodwin, "Empirical Findings on the Role of Structured Data, Unstructured Data, and their Combination for Automatic Clinical Phenotyping," *AMIA Summits on Translational Science Proceedings*, vol. 2021, pp. 445–454, May 2021, ISSN: 2153-4063.

[21] W. K. Tan, S. Hassanpour, P. J. Heagerty, *et al.*, "Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain," en, *Academic Radiology*, vol. 25, no. 11, pp. 1422–1432, Nov. 2018, ISSN: 1076-6332. DOI: 10.1016/j.acra.2018.03.008.

[22] Icahn School of Medicine at Mount Sinai, *Mount Sinai Data Warehouse — Scientific Computing and Data*, en-US, 2023. [Online]. Available: https://labs.icahn.mssm.edu/msdw/ (visited on 07/11/2023).

[23] J. Cohen, "A Coefficient of Agreement for Nominal Scales," en, *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, ISSN: 0013-1644, 1552-3888. DOI: 10.1177/001316446002000104.

[24] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

[25] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, *spaCy: Industrial-strength Natural Language Processing in Python*, original-date: 2014-07-03, 2020. DOI: 10.5281/zenodo.1212303. [Online]. Available: https://github.com/explosion/spaCy (visited on 06/10/2023).

[26] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," en, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327, 2019, Conference Name: Proceedings of the 18th BioNLP Workshop and Shared Task Place: Florence, Italy Publisher: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5034.

[27] H. Eyre, A. B. Chapman, K. S. Peterson, *et al.*, *Launching into clinical space with medspaCy: A new clinical text processing toolkit in Python*, arXiv:2106.07799 [cs], Jun. 2021. DOI: 10.48550/arXiv.2106.07799.

[28] B. van Aken, I. Trajanovska, A. Siu, M. Mayrdorfer, K. Budde, and A. Loeser, "Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?" In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, Online: Association for Computational Linguistics, Jun. 2021, pp. 35–40. DOI: 10.18653/v1/2021.nlpmc-1.5.

[29] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," eng, *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977, ISSN: 0006-341X.

[30] M. C. V. Lieke M Spekhorst and D. I. o. C. a. Colitis (ICC), "Performance of the Montreal classification for inflammatory bowel diseases," en, *World Journal of Gastroenterology*, vol. 20, no. 41, pp. 15 374–15 381, Nov. 2014, Publisher: Baishideng Publishing Group Inc. DOI: 10.3748/wjg.v20.i41.15374.

[31] K. Krishnaprasad, J. M. Andrews, I. C. Lawrance, *et al.*, "Inter-observer agreement for Crohn's disease sub-phenotypes using the Montreal Classification: How good are we? A multi-centre Australasian study," *Journal of Crohn's and Colitis*, vol. 6, no. 3, pp. 287–293, Apr. 2012, ISSN: 1873-9946. DOI: 10.1016/j.crohns.2011.08.016.

[32] S. Shrestha, O. Olén, C. Eriksson, *et al.*, "The use of ICD codes to identify IBD subtypes and phenotypes of the Montreal classification in the Swedish National Patient Register," eng, *Scandinavian Journal of Gastroenterology*, vol. 55, no. 4, pp. 430–435, Apr. 2020, ISSN: 1502-7708. DOI: 10.1080/00365521.2020.1740778.

[33] A. N. Kho, J. Yu, M. S. Bryan, *et al.*, "Privacy-Preserving Record Linkage to Identify Fragmented Electronic Medical Records in the All of Us Research Program," en, in *Machine Learning and Knowledge Discovery in Databases*, P. Cellier and K. Driessens, Eds., ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2020, pp. 79–87, ISBN: 978-3-030-43887-6. DOI: 10.1007/978-3-030-43887-6_7.

[34] H. Dong, M. Falis, W. Whiteley, *et al.*, "Automated clinical coding: What, why, and where we are?" en, *npj Digital Medicine*, vol. 5, no. 1, pp. 1–8, Oct. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-022-00705-7.

[35] X. Yang, A. Chen, N. PourNejatian, *et al.*, "A large language model for electronic health records," en, *npj Digital Medicine*, vol. 5, no. 1, pp. 1–9, Dec. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-022-00742-2.

[36] OpenAI, *GPT-4 Technical Report*, arXiv:2303.08774 [cs], Mar. 2023. DOI: 10.48550/arXiv.2303.08774.