Development of an EHR-Based Phenotyping Algorithm to Identify Recommendations for Bedrest and Activity Restriction in Pregnancy

Ananya Rajagopalan[1], Ari Klein, PhD[1], Heather Williams, MS[1], Emily Schriver, MS[1], Beth Pineles, MD, PhD[1], Danielle Mowery, PhD, MS, MS, FAMIA[1]
[1]University of Pennsylvania, Philadelphia, PA, USA

Abstract

Preterm birth is a leading cause of neonatal mortality and morbidity worldwide. Doctors frequently recommend physical activity restrictions to pregnant patients with the hopes of preventing preterm birth, even though this practice has been proven ineffective and sometimes harmful. While clinicians document their activity restriction recommendations in clinical notes, they can be difficult to parse due to their unstructured format, and the prevalence of these recommendations has not been characterized on a large scale. In this study, we develop a phenotyping algorithm using large language models (LLMs) to identify activity restriction recommendations in clinical notes from Penn Medicine. Using 200 randomly selected patient records from PennChart (Epic Clarity), three trained reviewers annotated notes at the sentence-level to create a reference standard. We evaluated the OpenAI gpt-oss-20b model using iterative prompt refinement to classify the presence and type of activity restriction. The model demonstrated strong off-the-shelf performance in early batches, with high precision and recall for certain restriction types (work, pelvic rest, and lifting in batch 1, bed rest in batch 2), though performance varied and favored precision over recall for overall restriction identification. These results support the use of LLM-based NLP methods to identify activity restriction recommendations and inform efforts to de-implement them nationally.

Motivation

Preterm birth (PTB), defined as delivery before 37 weeks of gestation, is a leading cause of neonatal mortality and morbidity worldwide[1], occurring in approximately half a million births per year. PTB remains a substantial public health challenge and accounts for approximately $25 billion in healthcare costs annually[2]. Despite extensive research on this topic, no population-scale intervention has successfully reduced either the overall incidence of PTB or the persistent disparities associated with it.

Of the numerous strategies have been attempted to prevent PTB and miscarriage, bedrest and activity restriction have been among the most frequently used, being recommended for over a century. However, there is strong evidence to demonstrate that this practice is ineffective and sometimes harmful[3]. Documented adverse effects include increased risk of venous thromboembolism, worsened mental health outcomes such as depression, loss of bone density, and significant financial and social burdens for patients and their families[4]. Moreover, studies indicate that activity restriction may actually increase the risk of PTB among individuals already at high risk, such as those with a shortened cervix[5,6]. Professional organizations including the American College of Obstetricians and Gynecologists (ACOG) and the Society for Maternal-Fetal Medicine (SMFM) explicitly recommend against the use of bedrest and activity restriction for the prevention of PTB[4,7].

Despite clear guideline recommendations, the current prevalence of bedrest and activity restriction and the clinical contexts in which they are prescribed remain poorly characterized. Existing evidence is limited to small studies, which suggest that these practices continue to be common. For example, a single-institution study conducted within a health system that actively discouraged activity restriction found that 37% of patients at high risk for PTB still received recommendations for restricted activity[8]. These findings suggest that the dissemination of guidelines and their adoption in local clinical settings is insufficient. However, up-to-date, large-scale data describing how often bedrest and activity restriction are recommended are lacking.

Solution

Recommendations for bedrest and activity restriction are typically documented in free-text clinical notes within the electronic health record (EHR), rather than as structured orders or diagnosis codes. While structured data such as ICD-10 codes and medication records have traditionally been used for clinical research, unstructured narrative text contains

substantial clinical detail that is increasingly being leveraged for research purposes. Given the volume and variability of clinical documentation, advanced computational approaches are necessary to extract meaningful information. Natural language processing (NLP) integrates techniques from linguistics, statistics, and machine learning to enable automated analysis and interpretation of text. Because clinicians use diverse language to describe activity recommendations, NLP methods are well suited to identify and categorize documentation of bedrest and activity restriction. Large language models (LLMs) in particular have emerged as a powerful tool for extracting clinically meaningful information from unstructured EHR data. Prior work has demonstrated that these transformer-based models can identify phenotypes and clinical concepts from narrative notes (such as strokes, cardiac arrests, thyroid dysfunction symptoms, and eosinophilic esophagitis) with performance comparable to or exceeding traditional rule-based or machine learning approaches, while requiring substantially less manual feature engineering and annotation effort[?,9–13]. These findings highlight the potential of LLMs to characterize understudied clinical practices using unstructured EHR data. **In this work, we aim to use LLMs to characterize the prevalence and nature of activity restriction recommendations using clinical notes from providers at Penn Medicine.**

Our approach leverages an open-source large language model (OpenAI gpt-oss-20b) to phenotype activity restriction recommendations from clinical notes. This task is especially well suited to LLM-based methods, as clinicians use highly variable and context-dependent language to describe restrictions, which limits the effectiveness of keyword-based or rule-driven approaches. Compared to manual chart review, LLMs offer a scalable and reproducible means of identifying both the presence and type of activity restriction recommendations across large patient populations, enabling systematic measurement of a practice that is otherwise difficult to quantify. One important drawback of using LLMs for this EHR-based phenotyping task is low sensitivity (recall) in detecting patient-reported activity restriction recommendations. This is a likely outcome, given the low prevalence of activity restriction recommendations at the sentence level compared to the total volume of sentences in the clinical note, but this will be important information to obtain.

Methods

Dataset: Using PennChart data from Epic Clarity, our team obtained a dataset of de-identified clinical notes from patients with high-risk pregnancies. A total of 9,703 patients experienced high risk pregnancy risks, among 92,119 OB visits among Penn patients between 2014 and 2023. To select relevant notes, patients were identified based on having an obstetric episode within the past 10 years with a documented estimated due date (EDD). Among these pregnancies, those with relevant ICD-10 diagnoses associated with preterm birth risk were selected if the diagnosis occurred between 280 days before and 14 days after the EDD. All eligible encounters within the OB episode (including inpatient, outpatient, procedural, and telemedicine visits) were then identified. Clinical notes from these encounters were extracted for analysis.

Approximately 200 patient records of the total cohort were randomly selected and their corresponding clinical notes were labeled at the sentence-level by three trained reviewers using the eHost annotation tool, documenting whether activity restriction was recommended and the category of restriction (**Figure 1**).

These notes were split into 20 batches that are sequentially annotated and adjudicated before I receive them for the LLM analysis (the annotation and adjudication is ongoing). Thus far, I have received and processed the first two batches, each of which contains 10 clinical notes, so the subsequent results and discussion in this report are based on this subset of data.

Sentence-Level Annotation Algorithm: Given the low prevalence of activity restriction recommendations at the sentence-level per note, we focused the primary phenotyping task to be the overall presence of a restriction recommendation, along with the type of restriction recommendation if there was one, leaving the detection of modifiers (diagnosis, restriction strength or duration) as an exploratory analysis for later. Therefore, the **three binary classification tasks explored in this work were**:

1. Does a given sentence contain an activity restriction? (0/1, detecting from positive and negative sentences)
2. Two restriction type-specific classification tasks:

(a) Does a given sentence contain an activity restriction of the particular type *X*? (0/1, detecting from positive and negative sentences)

(b) Given a sentence with an activity restriction, is it of type *X*? (0/1, detecting from only positive sentences)

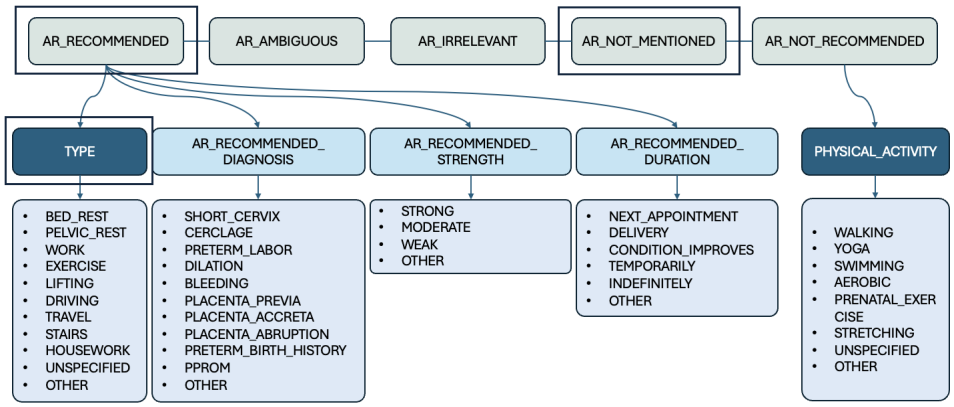

Figure 1: Annotation workflow of different classes and modifiers of activity restrictions. The annotation objective is to identify and classify activity restriction recommendations in clinical notes for pregnant individuals at high risk for preterm birth, including the type, duration, and strength of, and the diagnosis associated with, the recommendation. Boxed labels indicate the binary classification tasks that we used the LLMs for in this project.

Data Processing and LLM Workflow: In order to produce a phenotyping algorithm with the set of manually annotated clinical notes, I used Open AI's open weight model, gpt-oss-20b. To maintain patient confidentiality, the LLM analysis was conducted locally using Penn Medicine's HIPAA-compliant, Microsoft Azure Databricks tenant. Since the notes were manually annotated and adjudicated (disagreements between annotators resolved), I used the adjudicated annotations as a unified gold standard to compare the LLM performance against. **Figure 2** describes the workflow I followed to evaluate and iteratively improve the LLM's phenotyping abilities across subsequent batches of clinical notes.

I further processed the adjudicated annotations for each batch before designing, evaluating and refining my LLM prompt as described in **Figure 2**. In particular, I used the spans (character locations) for each annotation within a note to re-align the annotations with the corpus (original clinical note) to extract the full sentences that the annotations came from - since typically, the annotations themselves were highlighted over one or a few words within the sentence. I used RegEx to parse the file for \n\n+, which was how I determined individual sentences from the clinical note (the note had been processed using a sentence splitter previously, allowing me to use this simple logic to calculate the character offsets for each sentence). Once I identified the location and content for each annotated sentence (representing a positive for having an activity restriction recommendation), I added all sentences without annotations as negatives (controls), specifying their label as 'NOT_MENTIONED' in accordance with the annotations in **Figure 1**. I removed sentences with labels of 'NOT_RECOMMENDED', 'RECOMMENDED_STRENGTH', 'RECOMMENDED_DURATION', 'RECOMMENDED_DIAGNOSIS', or 'AMBIGUOUS' to maintain the classification task as a simple binary one between 'RECOMMENDED' and 'NOT_MENTIONED', as well as the presence of particular restriction types, as described in the Sentence-Level Annotation Algorithm section. Finally, I one-hot encoded the columns indicating the presence of an activity restriction or particular type of restriction, and collapsed all rows with duplicated sentences by aggregating (taking the max) for each column pertaining to an activity restriction. This ensured that there would be only one row per unique sentence in a note, with columns indicating different types of restrictions. I also made copies of the final, processed batch annotations with only positive sentences; I used this set for binary classification task 2b as described earlier.

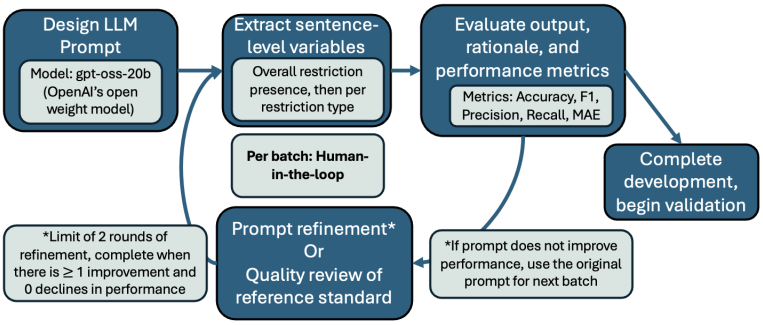

Figure 2: Specific workflow of iterating with LLMs across batches.

Results and Discussion

Figure 3 contains classification metrics for Batch 1 before and after refining the prompt (according to the heuristic in **Figure 2**), showing strong but varied performance across the different classification tasks with strength in identifying particular restriction types of pelvic rest, lifting, work, and exercise (especially in cases with just positive sentences). The model struggled to characterize sentences with unspecified restriction types, suggesting that the ambiguity around the definition of an unspecified restriction was a challenge for the LLM. **Figure 4** contains metrics for Batch 2 with the original prompt, showing similar results to Batch 1 in terms of having strength in detecting certain types of activity restrictions (bed rest), but weaker performance with other tasks.

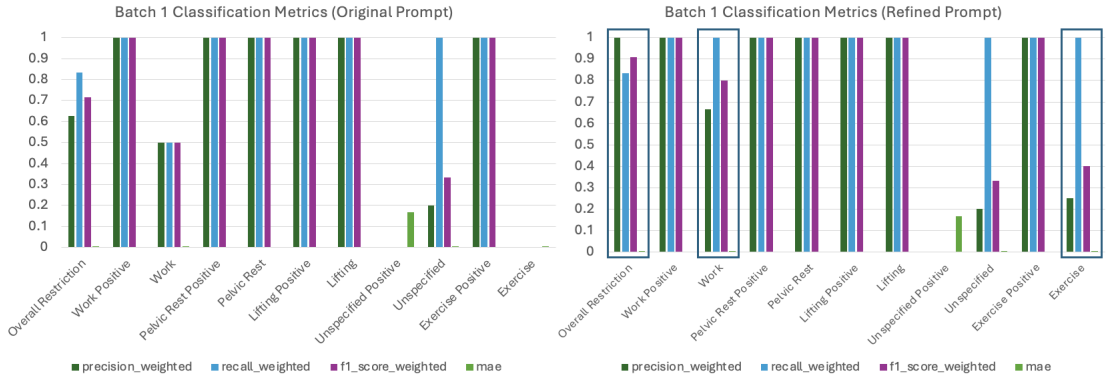

Figure 3: Batch 1 classification metrics, original and refined prompting. Boxed bars show the tasks for which refining the prompt improved performance (other task performances remained unchanged. I did not refine prompts for tasks where the LLM had perfect initial performance.

Each batch contained 10 notes. The time taken per classification task was approximately 1 second per sentence. The average time taken (seconds) for each classification task across batches is as follows: Batch 1: 1208.23 (Overall Restriction), 6.07 (Exercise Positive), 1152.79 (Exercise), 4.97 (Lifting Positive), 1189.02 (Lifting), 7.75 (Pelvic Rest Positive), 1204.34 (Pelvic Rest), 11.12 (Physical Unspecified Positive), 1358.79 (Physical Unspecified), 5.72 (Work Positive), and 1193.81 (Work); Batch 2: 1253.05 (Overall Restriction), 6.840 (Bed Rest Positive), 1168.53 (Bed Rest), 5.95 (Blunt Force Positive), 1281.26 (Blunt Force), 7.89 (Exercise Positive), 1198.83 (Exercise), 7.57 (Pelvic Rest Positive), and 1164.98 (Pelvic Rest).

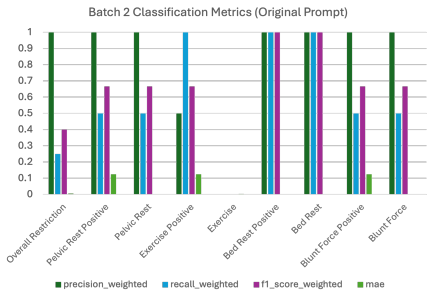

Figure 4: Batch 2 classification metrics.

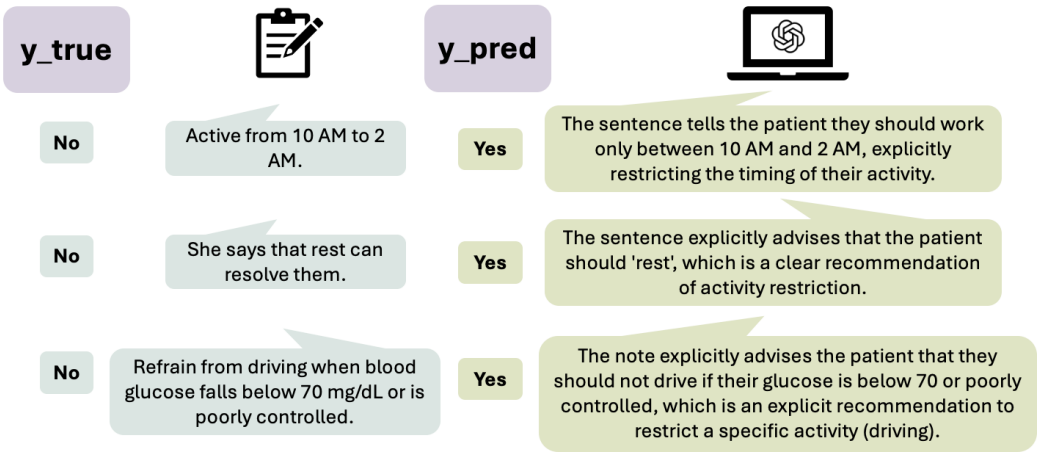

Figure 5: False positives for batch 1 overall classification task (presence of activity restriction per sentence, or not). Sentences displayed from the clinical note are synthetic sentences generated for the purposes of visualization.

I examined misclassified sentences for each classification task to better inform my prompt refinement strategy. **Figure 5** shows examples of the comparisons I was making for the overall restriction classification task in Batch 1. From the sentences, it was clear that the LLM was mistaking descriptions of patient state (e.g., saying they are active during certain hours of the day), as instructions from the doctor to the patient instead. Additionally, certain classifications (like telling the patient to refrain from driving conditional on a blood glucose value) was not obviously incorrect, but more nuanced and perhaps not what the annotators were looking for. In my later discussions with the team reviewing these findings, they mentioned that some of the LLM's misclassified sentences also came up in their adjudication discussions, highlighting the underlying ambiguity of some cases. When iterating my prompt, I chose to provide more context for the LLM on what *types* of restrictions I was looking for (what counted and what did not count). Additionally, for certain types of restrictions where the performance could have been improved, I provided more specific context on examples of that restriction (e.g., for the work restriction, I specified that "This is an exercise restriction and refers to exercise and workouts."). This improved the model's ability to identify certain types of restrictions, but in other cases, I found that providing too many (or too specific) instructions harmed performance marginally compared to simpler, more concise instructions. This is perhaps related to the trend that in Batch 2, there seemed to be consistently more false negatives than in Batch 1, suggesting that the refined prompts caused the LLM to be overly cautious in classifying an activity restriction. In future batches, I will need to balance the tradeoff between specificity and generalizability in designing prompts.

There are several possibilities in terms of both completing and extending this work. In the near term, next steps include completing the LLM evaluation for all remaining batches. We anticipate seeing incremental improvement (and eventually a plateau) in the performance of the LLM, once the prompt is refined to have better recall in identifying true positive cases of activity restriction recommendations. Additionally, I would like to evaluate the phenotyping abilities of other LLM architectures, specifically GPT o3 mini, GPT 5, and BERT. GPT 5 and o3 mini are of interest given their strong performance on phenotyping tasks from clinical notes in prior work by the team. BERT is a significantly smaller language model in terms of its parameter size, and whether its performance compares to larger models will be important to investigate.

These findings provide a foundation for understanding the prevalence of activity restriction recommendations in high-risk pregnancies, enabling actionable feedback to clinicians and health systems. By establishing a scalable method to measure a practice that lacks structured documentation, this work supports national efforts led by Dr. Pineles to de-implement activity restriction for preterm birth prevention in accordance with established guidelines. More broadly, accurate measurement of clinical practice is essential to improving pregnancy care and reducing harm from ineffective interventions. This study demonstrates the potential of LLMs to extract clinically meaningful signals from unstructured EHR data, supporting evidence-based pregnancy care and maternal health research.

References

1. Ohuma EO, Moller AB, Bradley E, Chakwera S, Hussain-Alkhateeb L, Lewin A, et al. National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. The Lancet. 2023 Oct;402(10409):1261-71.

2. Waitzman NJ, Jalali A, Grosse SD. Preterm birth lifetime costs in the United States in 2016: An update. Seminars in Perinatology. 2021 Apr;45(3):151390.

3. Saccone G, Della Corte L, Cuomo L, Reppuccia S, Murolo C, Napoli FD, et al. Activity restriction for women with arrested preterm labor: a randomized controlled trial. American Journal of Obstetrics & Gynecology MFM. 2023 Aug;5(8):100954.

4. Lauder J, Sciscione A, Biggio J, Osmundson S. Society for Maternal-Fetal Medicine Consult Series #50: The role of activity restriction in obstetric management. American Journal of Obstetrics and Gynecology. 2020 Aug;223(2):B2-B10.

5. Grobman WA, Gilbert SA, Iams JD, Spong CY, Saade G, Mercer BM, et al. Activity Restriction Among Women With a Short Cervix. Obstetrics & Gynecology. 2013 Jun;121(6):1181-6.

6. Levin HI, Sciscione A, Ananth CV, Drassinower D, Obican SG, Wapner RJ. Activity restriction and risk of preterm delivery. The Journal of Maternal-Fetal & Neonatal Medicine. 2018 Aug;31(16):2136-40.

7. Syed H, Slayman T, DuChene Thoma K. ACOG Committee Opinion No. 804: Physical Activity and Exercise During Pregnancy and the Postpartum Period. Obstetrics & Gynecology. 2021 Feb;137(2):375-6.

8. Bitar G, Sciscione A. The Compliance of Prescribed Activity Restriction in Women at High Risk for Preterm Birth. Am J Perinatol. 2022 Jan;39(01):054-60.

9. Yuan K, Yoon CH, Gu Q, Munby H, Walker AS, Zhu T, et al. Transformers and large language models are efficient feature extractors for electronic health record studies. Commun Med. 2025 Mar;5(1):83.

10. Hwang S, Reddy S, Wainwright K, Schriver E, Cappola A, Mowery D. Using Natural Language Processing to Extract and Classify Symptoms Among Patients with Thyroid Dysfunction. In: Bichel-Findlay J, Otero P, Scott P, Huesing E, editors. Studies in Health Technology and Informatics. IOS Press; 2024. .

11. Vurgun U, Kaviyarasu A, Hwang S, Batugo A, Thomas S, Tang B, et al.. Evaluating Large Language Models for Automatic Detection of In-Hospital Cardiac Arrest: Multi-Site Analysis of Clinical Notes. Health Informatics; 2025.

12. Ketchem CJ, Vurgun U, Wang A, Thomas S, Batugo A, Falk G, et al. S919 Large Language Models Combined With a Single Diagnostic Code Detect Eosinophilic Esophagitis in Electronic Health Records With High Accuracy. Am J Gastroenterol. 2025 Oct;120(10S2):S198-8.

13. Yang A, Kamien S, Davoudi A, Hwang S, Gandhi M, Urbanowicz R, et al. Relation Detection to Identify Stroke Assertions from Clinical Notes Using Natural Language Processing. In: Bichel-Findlay J, Otero P, Scott P, Huesing E, editors. Studies in Health Technology and Informatics. IOS Press; 2024. .