An Iterative Annotation Pipeline for Building Clinical Datasets and Training Information Extraction Models: The PREPARE Project

Erik Calcina[1,2,*], Erik Novak[1,2], Lucia Pepa[3], Maria Gabriella Ceravolo[3] and the PREPARE Project Group[3]

[1]**Department for Artificial Intelligence, Jozef Stefan Institute, Ljubljana, Slovenia**

[2]**Jozef Stefan International Postgraduate School, Ljubljana, Slovenia**

[3]**Marche Polytechnic University, Marche, Italy**

[3]*The following individuals are members of the* PREPARE Project *Group. Their contributions to the PREPARE project enabled this work:* Aghayev E., Alepopoulos V., Banfi G., Burger H., Calcina E., Capecci M., Ceravolo M-G, Cordani C., Ferrario I., Georgiadis Ch., Giannios G., Goker F., Gramatikopoulou M., Herrmann Ch., Hoogendam L., Janssen E., Kiekens C., Knoop J., Kompatsiaris I., Kotsis V., Kuret Z., Matjačić Z. Mpaltadros L., Negrini A., Negrini S., Nikolopoulos S., Novak E., Patias G-R, Patias P., Pennestrì F., Poli A., Roustanis Th., Selles R., Staal B., Tsaopoulos D., Tartaglia G., Vidmar G., Zaharatos H.

*Corresponding author: erik.calcina@ijs.si

Background

Electronic health records (EHRs) are a primary source of data for clinical care and biomedical research. However, much of the clinically relevant information is stored in unstructured free-text clinical notes containing abbreviations and inconsistent terminology, making reliable and scalable information extraction challenging. Accurate transformation of such data into structured formats is essential to support downstream analysis and clinical decision making. The challenge is particularly evident in specialized clinical domains, where linguistic variability limits the effectiveness and scalability of manual abstraction and simple rule-based approaches.

The EU-funded PREPARE[1] project aims to improve rehabilitation care for patients with chronic noncommunicable diseases through the development of data-driven prediction and stratification tools. The project is built around the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). While structured data can be directly mapped to the OMOP CDM, information in unstructured clinical notes must first be extracted before integration.

In this abstract, we extend our previous work[2] by introducing a new dataset of narrative data from individuals with neurodegenerative disorders and by presenting an improved information extraction approach. The method first constructs a high-quality annotated dataset using an iterative annotation pipeline within the PREPARE project, which is then used to support downstream model training and evaluation. Our goal is to enable fast, precise, low-resource extraction that runs efficiently on standard laptop hardware. While the extraction can be highly precise, clinician review is still required to confirm the correctness of the extracted information. We present both the methodology and results demonstrating the feasibility of this approach.

Methods

Manual extraction of clinical data from EHRs is inefficient and resource intensive, motivating the use of Named Entity Recognition (NER) models for automated information extraction. High-precision NER models require well-annotated training data, which is often unavailable for medical records. To address this limitation, we developed an iterative pipeline to transform unstructured clinical notes

into an annotated dataset suitable for model training. The pipeline focuses on dataset construction, with extraction models trained after annotation is finalized. This dataset was subsequently divided into train and test part and used to evaluate multiple large language model (LLM)–based and GLiNER-based extraction models.

***Data annotation proc*ess:** To construct the annotated dataset, we designed an iterative information extraction workflow based on LLMs; in our case, we used **LLaMA 3.1 Instruct**[3]. The pipeline combines zero-shot prompting, human annotation, and model fine-tuning. In the initial phase, zero-shot extraction is used to generate baseline predictions from unstructured clinical notes, which serve as support during the annotation process. These outputs are reviewed and corrected by human annotators using the Label Studio[4] platform and are then used to fine-tune the extraction model. The fine-tuned model is applied to larger set of clinical notes, and their predictions are reintroduced into subsequent annotation rounds. This cycle is repeated to progressively improve annotation quality and extraction accuracy, resulting in a robust annotated dataset.

Large Language Model fine-tuning is performed using parameter-efficient training techniques, in which only lightweight adapter modules are updated rather than the full model parameters. This approach substantially reduces memory requirements, limits catastrophic forgetting, and shortens training time. To further improve computational efficiency, all models are quantized to 4-bit precision during fine-tuning. Training is conducted in a supervised manner and is restricted to the model-generated outputs. Tokens that do not correspond to target labels, such as system prompts and input context, are masked during loss computation. This design ensures that the model is optimized specifically for producing the expected JSON-formatted label outputs, rather than learning the structure of the prompts or the input text itself.

***GLiNER*[5]** was fine-tuned following the training procedure recommended by the authors. The pre-trained model was adapted to the target extraction task by updating its parameters using supervised training, without modifications to the underlying architecture. This approach leverages GLiNER's span-based formulation for named entity recognition while remaining computationally efficient and suitable for execution on limited hardware resources.

Results

In the initial iteration, manual annotation required over two minutes per clinical note (Table 1). With successive rounds of fine-tuning and model-assisted annotation, annotation time decreased to under one minute per note and stabilized at approximately 45 s in later iterations.

Iteration	Training state of model	Number of annotated notes	Average time per note
1	Zero-shot prompting	50	2 min 43 s
2	Fine-tuned on 50 notes	50	1 min 4 s
3	Fine-tuned on 100 notes	100	50 s
4	Fine-tuned on 200 notes	143	45 s

Table 1. Average annotation time per clinical note across successive iterative refinement rounds, showing reduced annotator effort as model assistance and fine-tuning increased.

Several large language models and GLiNER variants were evaluated, with only the best-performing model from each group reported (Table 2). Performance was assessed using three F1 metrics. *Exact F1* measures cases where the entity span must exactly match the annotation. *Relaxed F1* considers predictions correct when the extracted span is a sub-span of the annotated entity. *Overlap F1* measures any overlap between predicted and annotated spans, independent of span completeness. For a prediction to be counted as correct under any variants, the label must also be correct.

In Table 2 we report the best-performing LLM and GLiNER model, which achieve comparable results across the evaluated metrics. Despite this similarity in performance, practical deployment constraints differ substantially. GLiNER models can be executed efficiently on low-resource machines, such as standard laptops, whereas LLM-based approaches require GPU acceleration for inference, limiting their deployment in resource-constrained clinical environments.

Model	Epochs	Exact F1	Relaxed F1	Overlap F1
gemma-3-12b-it[6]	6	0.7887	0.8183	0.8972
gliner_large-v2.1[7]	10	0.7776	0.7992	0.9485

Table 2. Performance of the best-performing large language model and GLiNER model evaluated using Exact F1, Relaxed F1, and Overlap F1 metrics.

Conclusion

This research presents an iterative methodology for efficiently constructing high-quality annotated datasets from unstructured clinical notes. The proposed pipeline combines zero-shot prompting, human-in-the-loop annotation, and parameter-efficient model fine-tuning to progressively reduce annotation effort while improving annotation consistency. The resulting dataset is then used to train and evaluate multiple extraction models, including LLMs and GLiNER, for downstream clinical information extraction.

Results demonstrate that model-assisted annotation substantially accelerates dataset creation, with the largest efficiency gains observed in early refinement iterations. Using the final annotated dataset, both model types achieved comparable extraction accuracy across multiple matching criteria, while differing in deployment requirements.

References

1. PREPARE: Personalised predictive rehabilitation for patients with chronic noncommunicable diseases. Horizon Europe research and innovation programme; Grant Agreement No. 101080288.
2. Calcina E, Meterc T, Shpanko N, Spindari L, Lindner E, Hoogendam L, Slijper H, Selles R, Novak E. An iterative LLM-based pipeline for extracting clinical data from medical records. In: *Proceedings of the OHDSI Europe Symposium 2025*; 2025.
3. AI@Meta. Llama 3 model card [Internet]. 2024 [cited 2026 Jan 30]. Available from: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
4. Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. Label Studio: Data labeling software [Internet]. 2020–2025 [cited 2026 Jan 30]. Available from: https://github.com/HumanSignal/label-studio
5. Stepanov I, Shtopko M. GLiNER multi-task: Generalist lightweight model for various information extraction tasks. arXiv:2406.12925 [Internet] [cited 2026 Jan 30]. Available from: https://arxiv.org/abs/2406.12925
6. Google. Gemma-3-12b-it [Internet]. Hugging Face; 2024 [cited 2026 Jan 30]. Available from: https://huggingface.co/google/gemma-3-12b-it

7. Urchade. gliner_large-v2.1 [Internet]. Hugging Face; 2024 [cited 2026 Jan 30]. Available from: https://huggingface.co/urchade/gliner_large-v2.1

Acknowledgments

PREPARE is funded by the European Union. UK participants are supported by UKRI grant number 10086219 (Trilateral Research). Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01