Rare-Xtract: A Hybrid Pipeline for Generating Real-World Data Products From Unstructured Electronic Medical Records (EMRs) of Rare Diseases Leveraging Natural Language Processing (NLP) and Keyword Extraction Techniques

RWD118

James A[1], Has C[2], Palucci O[2], Mertz E[2], Rentsch I[2]
[1]Centogene GmbH

Introduction/Objective

Real-world evidence (RWE) plays an essential role in deepening our understanding of disease progression, treatment outcomes, and the natural history[1] of rare diseases. Despite its significance, much of the available RWE, including electronic medical records (EMRs) and physician notes, exists in an unstructured format[2,3] which presents significant challenges for data extraction, interpretation, and analysis. To address this issue, we developed RARE-XTRACT, a sophisticated hybrid pipeline engineered to convert unstructured real-world data into curated, research-ready datasets by employing a combination of rule-based algorithms and fine-tuned large language models (LLMs)[4]. In this study, we illustrate the utility of RARE-XTRACT in processing consented unstructured data from patients diagnosed with Phenylketonuria (PKU) over a decade (2013–2023), encompassing data collected at CENTOGENE GmbH (CNTG) as well as from external healthcare providers.

Method

The RARE-XTRACT pipeline (Figure 1) operates in two stages. First, PDFs are classified into broad categories such as Physical Examination, Biochemical Data, Medical Notes, MRI, and Biopsy mentions. Each category has predefined subcategories (e.g., Physical Examination includes height, BMI, weight, blood pressure). A medical dictionary was used to map subcategory terms, and the classification tool labels each PDF page in under a minute.

In the second stage, external EHR records were analyzed, focusing on extracting drug, diet, and longitudinal biochemical data for PKU patients. Fine-tuned LLM models, including a Medical-NER model based on deBERTa[5] (trained on 41 medical entities from PubMED), were applied for entity recognition. Drug terms were identified and validated using a custom Drug NER library[6], and dose/unit information was extracted with in-house tools. Diet data was extracted using Spacy's PhraseMatcher[7], which matched predefined diet terms with EHR records. Subsequently, all extracted values and units underwent manual validation, and cleaning. Finally, these were systematically consolidated into a CSV file for further analysis.

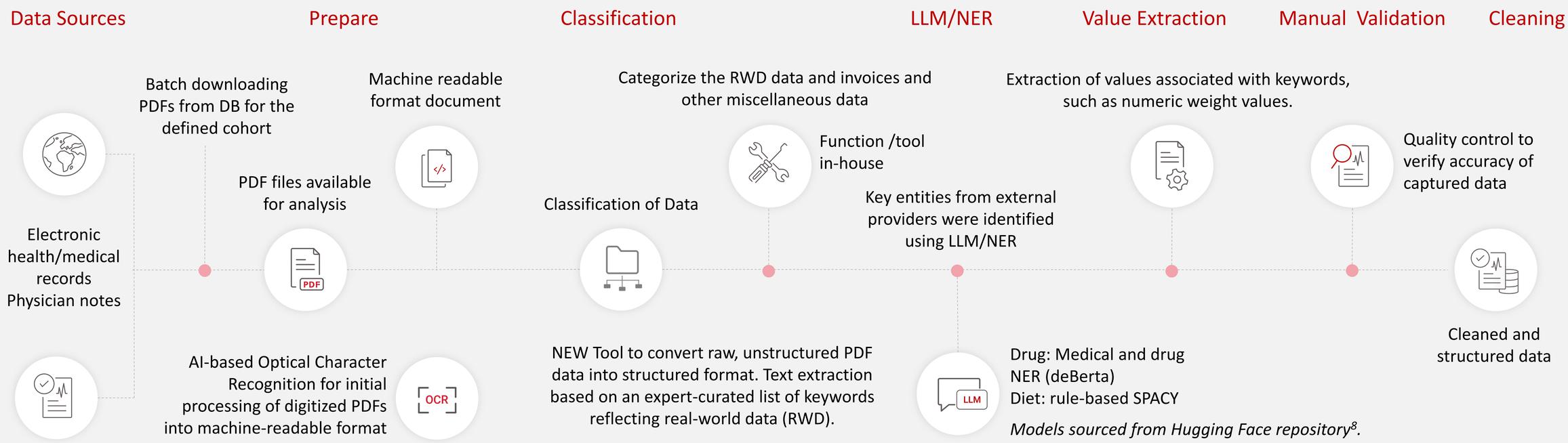

Figure 1: Developed cutting-edge generative AI-based entity recognition framework
The tool takes less than a minute to scan through a PDF document with under 100 pages with a precision of 86%

Results

We processed 1,082 unstructured PDF files containing clinical data from 255 PKU consented patients at CNTG (2013–2023) from multiple countries. The dataset included invoices, diagnosis reports, EHRs, EMRs, and physician notes. Besides that, we also processed structured text-based data from external providers for 200 PKU patients. Using a regular expression-based approach with manually curated keyword lists, we classified relevant data, reducing manual effort with each PDF processed in under a minute. Fine-tuned LLM models (Medical-NER) reached an average accuracy of 81.5% (Figure 2), while a rule-based model (SpaCy PhraseMatcher) for dietary extraction achieved 75% accuracy. The classification achieved an overall precision of 86%. Accuracy across broader categories was: 94% for follow-up notes, 92% for anamnesis, 78% for progress, 76% for history, and 53% for biochemical data and image files (Figure 3). Finally, the value extraction tasks averaged 61% accuracy, for the key metrics: Phenylalanine (75%), dietary values (76%), and Tyrosine (31%).

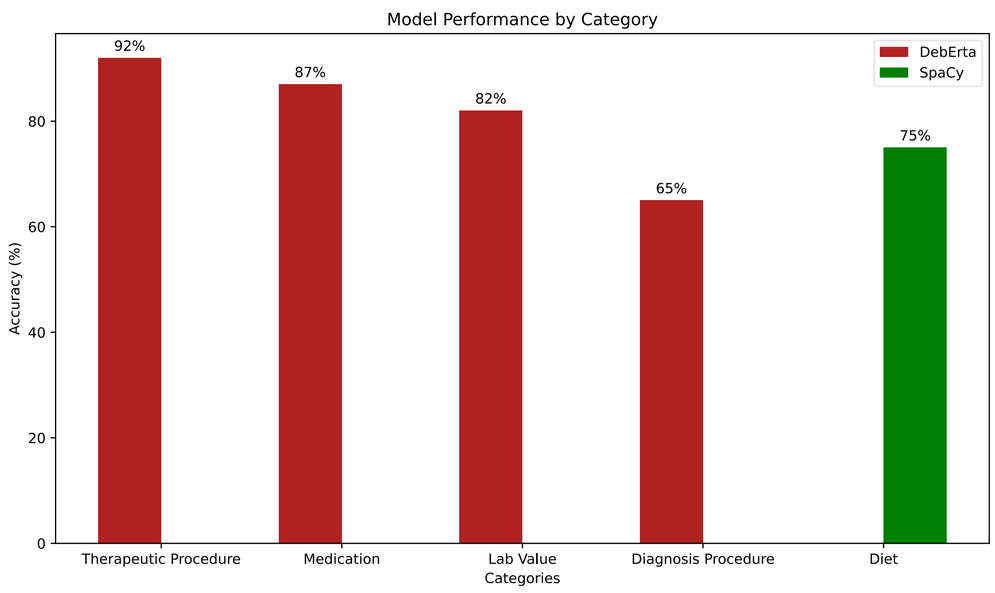

Figure 2: The accuracy of deBERTa based medical name entity recognition model for entities, therapeutic procedure, diagnosis, medication, and lab values.

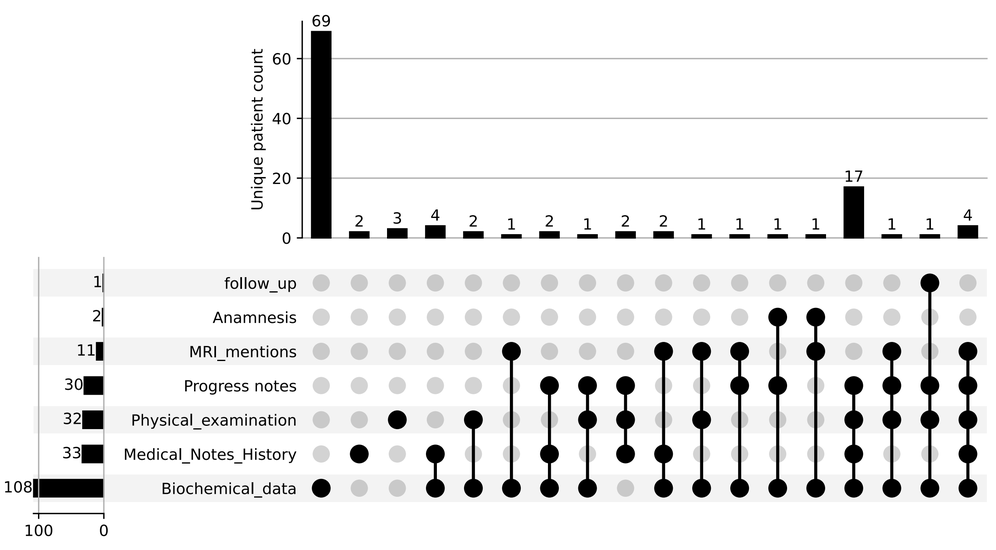

Figure 3: The upset plot depicting the broad category data with the number of patients per broad category.

Conclusion

The RARE-XTRACT pipeline represents a significant advancement in processing and transforming unstructured RWD for rare diseases. By integrating rule-based approaches with fine-tuned large language models, the tool automates the extraction and classification of critical clinical information with high precision and accuracy. This approach not only accelerates data processing but also enhances the quality and utility of RWD for rare disease research. RWD can be used in statistical models, where inferential approaches are essential for uncovering causal relationships, validating hypotheses, and generating regulatory-grade real-world evidence (RWE) to inform policymakers and regulators just as in controlled trial settings[9]. The demonstrated effectiveness of RARE-XTRACT in PKU highlights its broader potential to enable more informed insights into disease progression and patient management.

References

1. *Liu J, Barrett JS, Leonardi ET, Lee L, Roychoudhury S, Chen Y, Trifillis P. Natural History and Real-World Data in Rare Diseases: Applications, Limitations, and Future Perspectives. J Clin Pharmacol. 2022 Dec;62 Suppl 2(Suppl 2):S38-S55. doi: 10.1002/jcph.2134. PMID: 36461748; PMCID: PMC10107901.*
2. *Yang, X., Chen, A., PourNejatian, N. et al. A large language model for electronic health records. npj Digit. Med. 5, 194 (2022). https://doi.org/10.1038/s41746-022-00742-2*
3. *Liu, F., Panagiotakos, D. Correction: Real world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 23, 109 (2023). https://doi.org/10.1186/s12874-023-01937-1*
4. *Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., & Feder, A. (2023). LLMs Accelerate Annotation for Medical Information Extraction. Proceedings of Machine Learning Research, 225.*
5. *blaze999/Medical-NER · Hugging Face*
6. *Wood, T.A., Drug Named Entity Recognition [Computer software], Version 2.0.0, accessed at https://fastdatascience.com/drug-named-entity-recognition-python-library, Fast Data Science Ltd (2024)*
7. *PhraseMatcher · spaCy API Documentation*
8. *T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, HuggingFace's Transformers: State-of-the-art Natural Language Processing, CoRR. abs/1910.03771 (2019).*
9. *Liu, F., Panagiotakos, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 22, 287 (2022). https://doi.org/10.1186/s12874-022-01768-6*

CENTOGENE
THE RARE DISEASE COMPANY

For more Information
www.centogene.com