Integrating advanced **NLP techniques** with the **OMOP CDM** enhances the accuracy and completeness of **hematologic oncology data**

Enhancing Hematologic Oncology Data Capture in the OMOP CDM: Methodological Advances and Challenges

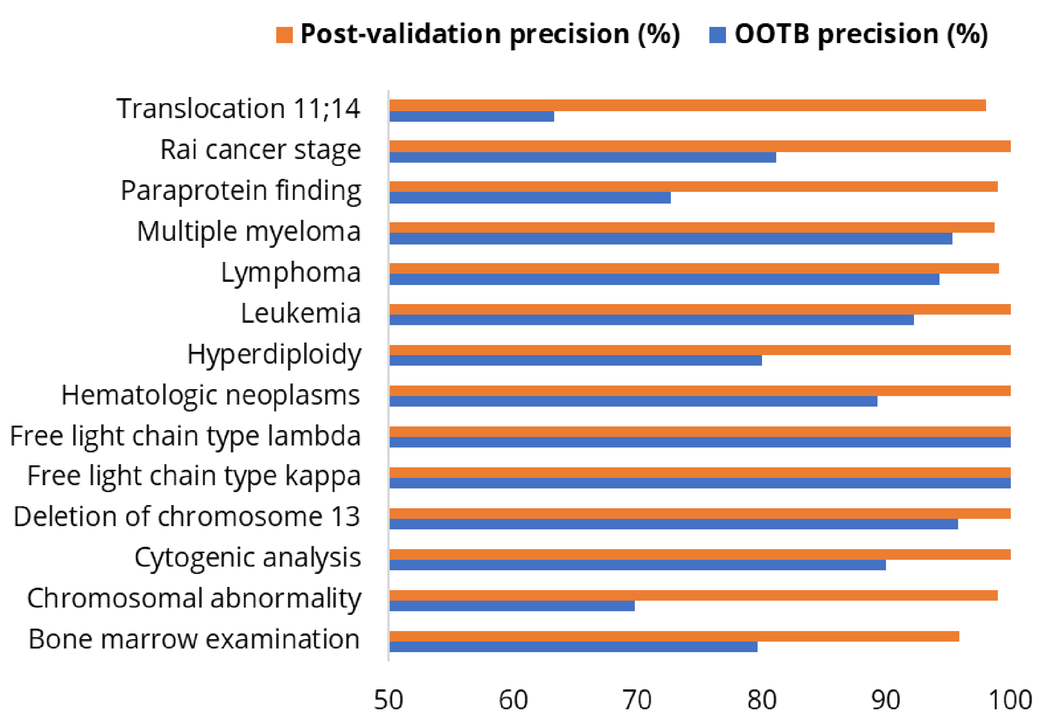
Background: Accurate data capture in hematologic oncology is crucial for understanding disease progression and outcomes, but traditional methods often miss complex variables. This study applies a natural language processing (NLP) pipeline within the OMOP CDM to enhance the capture of key variables in diseases like chronic lymphocytic leukemia (CLL) and multiple myeloma (MM), improving the quality and depth of oncology research data.

NLP Pipeline: We applied multilingual transformer-based models, trained on in-house curated data, for concept recognition, normalization and attribute extraction such as negation and temporality.

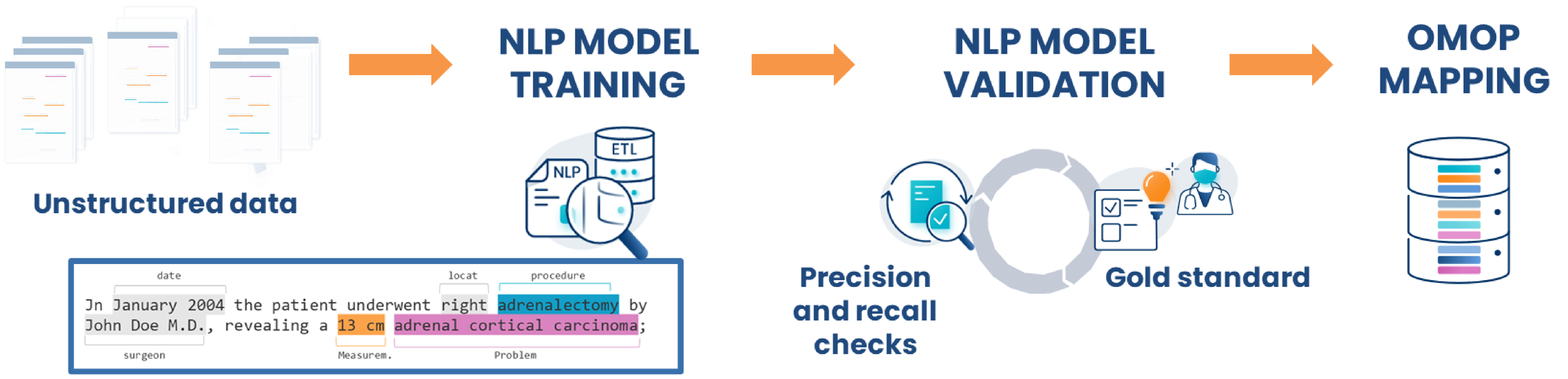
Finetuning: Out-of-the-box (OOTB) precision metrics show clear improvement upon finetuning during validation, based on expert feedback, leading to enhanced accuracy after refinement.

De patiënt heeft geen geschiedenis van diabetes in de familie, maar vertoont symptomen van hartfalen.
In 2011 is er een EGFR -positieve NSLC
vastgesteld, waarvoor behandeling opgestart met Pembrolizumab sinds mei 2011.
The patient has diabetes and is showing early signs of hypertension. Last year, she tested positive for rheumatoid arthritis
Der Patient hat eine Vorgeschichte von Asthma und Pankreaskarzinom.
La patient n'a pas d'antécédents de cancer, mais il présente des symptômes d 'hypertension.

phrase	cui	name	negation	dates	experiencer	uncertainty	temporality
diabetes	C0011849	Diabetes Mellitus	☑	2024-07-12	other	☐	historical
hartfalen	C0018801	Heart failure	☐	2024-07-12	patient	☐	recent
EGFR -positieve NSLC	C5770046	Primary epidermal gr	☐	2011-12-31	patient	☐	recent
Pembrolizumab	C3658706	pembrolizumab	☐	2011-05-31	patient	☐	recent
diabetes	C0011849	Diabetes Mellitus	☐	2024-07-12	patient	☐	recent
hypertension	C0020538	Hypertensive disease	☐	2024-07-12	patient	☑	recent
rheumatoid arthritis	C0003873	Rheumatoid Arthritis	☐	2023-12-31	patient	☐	historical
Asthma	C0004096	Asthma	☐	2024-07-12	patient	☐	historical
Pankreaskarzinom	C0346647	Malignant neoplasm	☐	2024-07-12	patient	☐	historical
cancer	C0006826	Malignant Neoplasm	☑	2024-07-12	patient	☐	recent
hypertension	C0020538	Hypertensive disease	☐	2024-07-12	patient	☐	recent



Methods



Conclusion: We demonstrate the effectiveness of integrating an NLP pipeline with structured data mining in OMOP CDM databases to improve data capture in hematologic oncology. High precision, recall, and F1 scores validate the reliability of this approach, which enhances the quality of datasets and supports large-scale, multicenter research. Our findings highlight the potential of NLP to significantly improve data management and insights in cancer research.

Clara L. Oeste[1], Iege Bassez[1], Jana Van Canneyt[1], Alina Kramchaninova[1], Lucas Sterckx[1], Narges Farokhshad[1], Shahbaz Pervaiz[1], Geert Van Gorp[1], Dries Hens[1]
[1]LynxCare Inc., Leuven, Belgium.

LYNXCARE