Enhancing Data Quality in Health Research: Performance Insights of a Clinical NLP Pipeline for Diverse Medical Domains

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

[1]LynxCare Inc., Leuven, Belgium.

Booth 1126 | MSR10

OBJECTIVES

This study presents performance insights into our clinical **Natural Language Processing (NLP) pipeline**. We have developed **multilingual transformer-based models**, trained on in-house curated data, for concept recognition, normalization and attribute extraction such as negation and temporality **(Figure 1).**

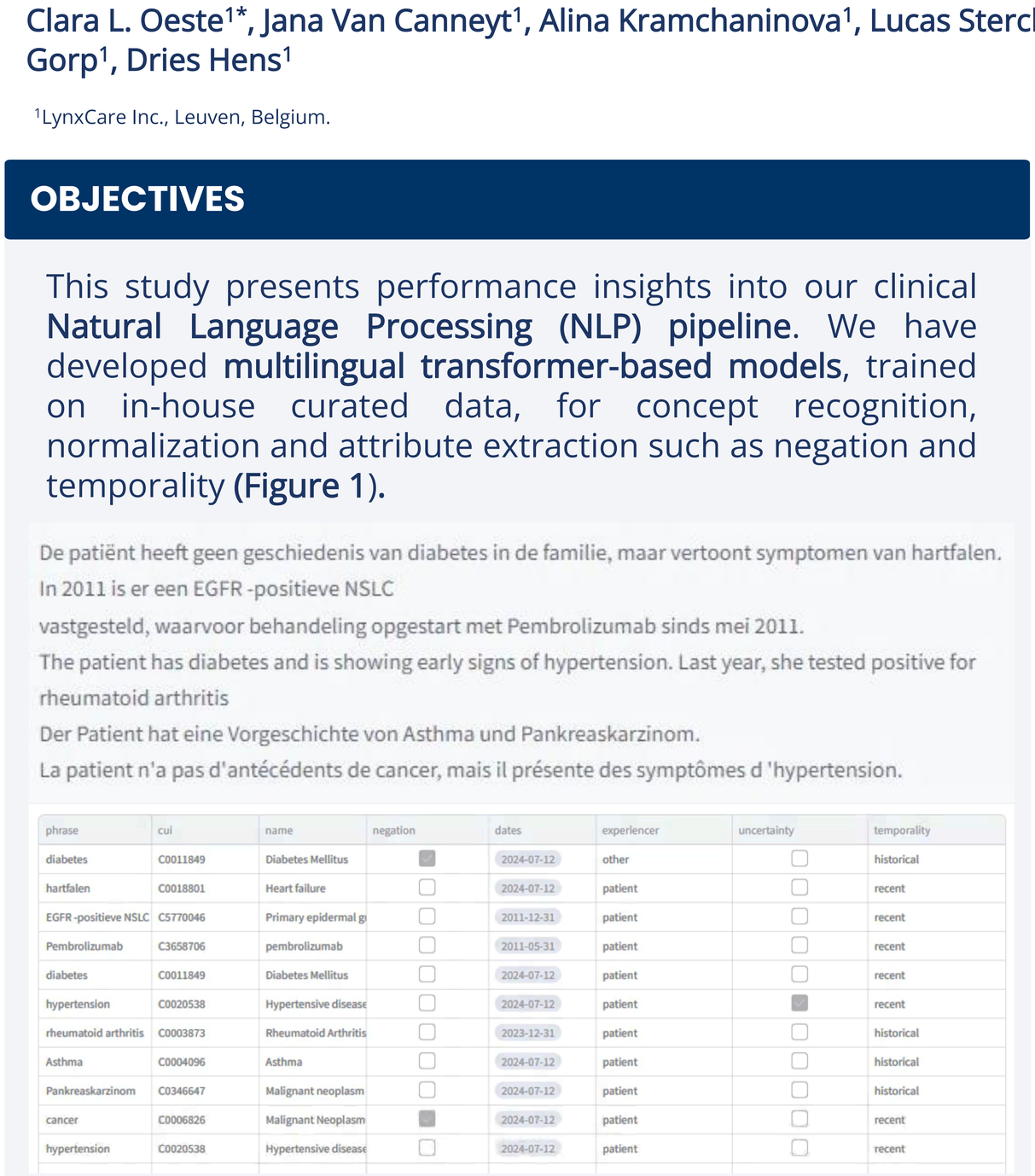


Figure 1. Example of multilingual NLP pipeline concept extraction and mapping

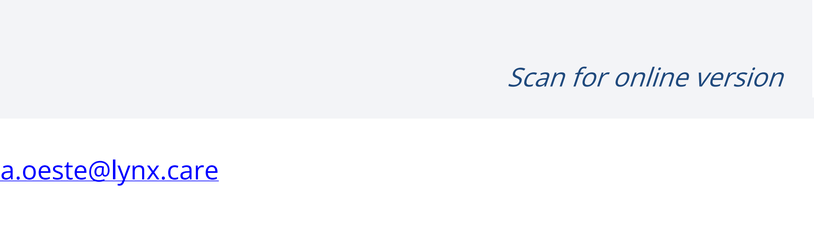
We **compare here the out-of-the-box (OOTB) precision and recall to post-validation metrics.**
We aim to:

- highlight the initial effectiveness of OOTB metrics, and
- demonstrate the crucial role of validation by annotators and physicians in boosting performance across diverse medical domains.

RESULTS

OOTB vs post-validation metrics were assessed for:

- **Broad-scope terms:** 58 data points and 7,207 records.
- **Oncology terms:** 193 data points and 13,215 records.
- **Cardiology terms:** 62 data points and 2,686 records.



Out-of-the-box NLP pipeline — Post-validation NLP pipeline

Term type	Precision OOTB	Recall OOTB	Precision post-validation	Recall post-validation
Broad	86.3%	82.1%	96.6%	96.3%
Oncology	82.3%	79.1%	96.8%	94.3%
Cardiology	88.6%	79.4%	96.6%	95.6%

Scan for online version

METHODS

The NLP pipeline processes unstructured data from **electronic health records (EHRs) of hospitals** within our network, focusing on broad-scope data points and specific therapeutic areas (TAs) to generate **OMOP-CDM databases** that also include structured data sources (**Figure 2**).

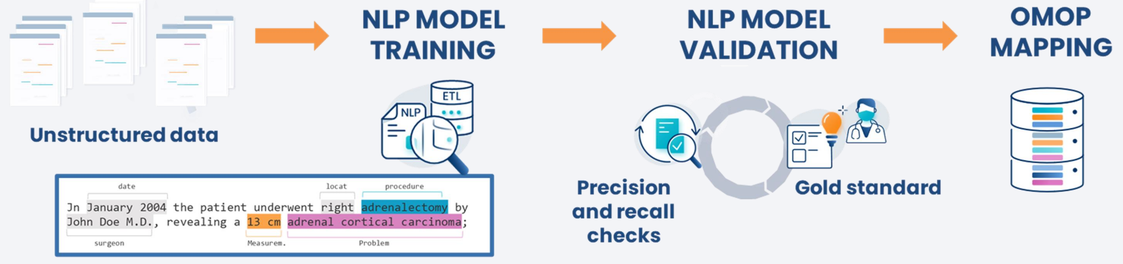


Figure 2. Unstructured data processing by NLP model training, validation, and mapping.

- Initial OOTB precision and recall were calculated for **312 data points in over 23,000 records.**
- Subsequently, **physicians reviewed** data point hierarchy and relevance, and validation against a human-generated gold standard (**Figure 3**).
- **Feedback is leveraged** to enhance the NLP pipeline regarding e.g., retraining of named entity recognition, entity linking, date detection and attribute detection components
- **Post-validation metrics** that were assessed:
  - **Precision** (true positives among all detected)
  - **Recall** (true positives among all actual data points)

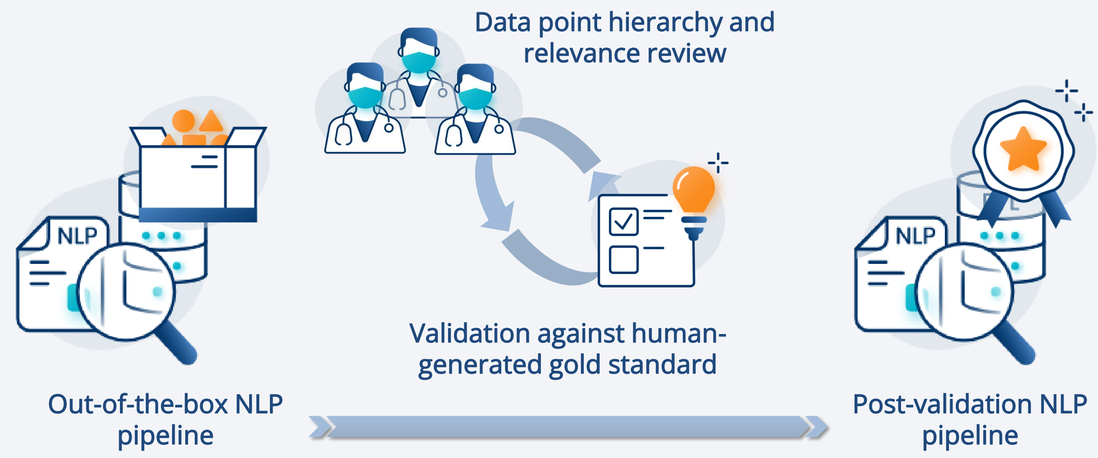


Figure 3. Validation feedback loop for NLP pipeline enhancement

CONCLUSIONS

- Our clinical **NLP pipeline significantly enhances real-world data (RWD) quality and integrity** by enriching OMOP-CDM datasets with unstructured data.
- Initial out-of-the-box (OOTB) metrics demonstrate promising results, with subsequent validation across diverse medical domains validating its effectiveness for continuous data enrichment.
- Its successful implementation in studies leading to peer-reviewed scientific manuscripts underscores its role in **supporting large-scale, cross-institutional research initiatives**, contributing to evidence-based medical insights.

Take-home message:
Validation by annotators and physicians ensures comprehensive quality assurance and enhances NLP pipeline performance.

*Contact: clara.oeste@lynx.care