Entity Linking Error Analysis Using NLP in Multilingual Clinical Narratives

Narges Farokhshad[1], Alina Kramchaninova[2], Stephanie Vandeput[1], Lucas Sterckx[1], Clara L. Oeste[1]

[1] LynxCare Clinical Informatics, Leuven, Belgium [2] LynxCare Clinical Informatics, Leuven, Belgium (formerly)

LYNXCARE INTRODUCTION

- **Healthcare Technology** — founded 2015, Belgium
- At **LynxCare**, we unlock clinical data by transforming unstructured medical narratives into **structured research-ready datasets** using AI and NLP.
- Built on the **OMOP Common Data Model** — enabling interoperability across hospital networks

50+ Clinical OMOP-CDM Databases

20+ Hospital Partners across Europe

55+ Peer-reviewed Publications (Manuscripts and abstracts)

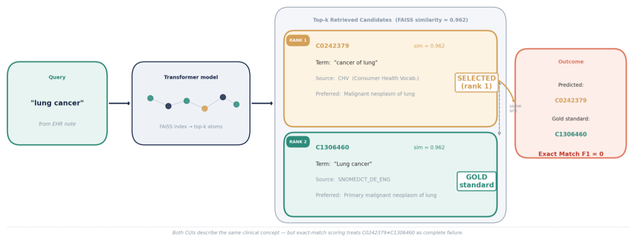
Key Clinical NLP Use Case at LynxCare
The LynxCare NLP pipeline extracts key medical concepts from clinical text and links them to standardized codes in medical ontologies (**UMLS**). This is the heart of ***clinical concept normalization***, transforming raw clinical language into structured, research-ready data.

BACKGROUND

Entity Linking (EL) maps clinical concepts from unstructured text, (e.g., EHR) to entries in structured medical ontologies like ***UMLS***.

Clinical Text → NER Extract Concepts → Entity Linking Assign CUIs → UMLS CUI

- In clinical NLP, EL underpins the extraction of real-world evidence from clinical narratives.
- Standard evaluation metrics treat all predictions as binary: **exact match or failure**.
- This misses clinically valid predictions.



OBJECTIVE

To computationally demonstrate the necessity of **alternative, relaxed evaluation methods** for clinical EL, given:

- Dynamic nature of clinical terminology
- Multilingual complexity across EHR systems
- Semantic richness missed by exact-match metrics
- Incomplete and evolving reference ontologies

METHOD – EVALUATION PIPELINE

1. **Run multilingual transformer-based EL model**
   Proprietary model generates predictions across all three test sets
2. **Separate exact matches**
   Non-matches flagged for human expert review
3. **Expert classification**
   Predictions labelled as CM · PIM · CIM · IM
4. **Compute extended metrics**
   F1 · Semantic Similarity · ACC@1/5/10 · MRR@10

ENTITY LINKING: ERROR TYPE CATEGORIES

✅ CORRECT MATCH (CM)

*"...bilateral knees pain with associated **swelling**."*

Concept: **swelling**
Gold: C0013604 — Edema
Predicted: C0038999 — Swelling

↳ *Different CUIs, clinically synonymous — both correctly describe the condition.*

⚠️ PARTIALLY INCORRECT (PIM)

*"...late **familial hyperinsulinemic** hypoglycemia due to a well-known mutation..."*

Concept: **familial hyperinsulinemic hypoglycemia**
Gold: C1847555 — Hyperinsulinemic hypoglycemia, familial, 6
Predicted: C3888018 — Congenital Hyperinsulinism

↳ *Related condition but at a broader level of granularity — informative but imprecise.*

🔶 CONTEXTUALLY INCORRECT (CIM)

*"...patient re-presented with a severe flare of **PV** and a recurrent deep vein thrombosis..."*

Concept: **PV**
Gold: C0030809 — Pemphigus Vulgaris
Predicted: C0030840 — Penicillin V

↳ *'PV' is a valid abbreviation for both — requires sentence context to disambiguate.*

❌ INCORRECT Match (IM)

*"...présence d'une **masse** latéroutérine droite mal systématisée..."*

Concept: **masse (FR: mass)**
Gold: C0577559 — Mass of body structure
Predicted: C0728811 — Masse brand of topical emollient

↳ *Completely wrong semantic domain: anatomical mass vs. pharmaceutical brand name.*

EVALUATION DATA

Dutch 1,000 concepts

French 529 concepts

English 445 concepts

- **EN/FR:** E3C public corpus test sets. (**Domain**: Case reports)
- **NL:** Proprietary dataset. (**Domain**: clinical EHR)
- **UMLS 2023AB** as reference ontology (**69.6% English** concepts, **2.9% other** languages)

F1
Harmonic mean of precision & recall. Binary: exact match only.

ACC@k
Correct concept appears in top-k predictions (k=1, 5, 10). Tolerates valid near-misses.

Semantic Similarity (SS)
Mean cosine similarity between predicted and reference embeddings.

MRR@10
Mean Reciprocal Rank — rewards higher placement of correct concept in ranked list.

RESULTS

Exact Match Metrics Underestimate True Performance

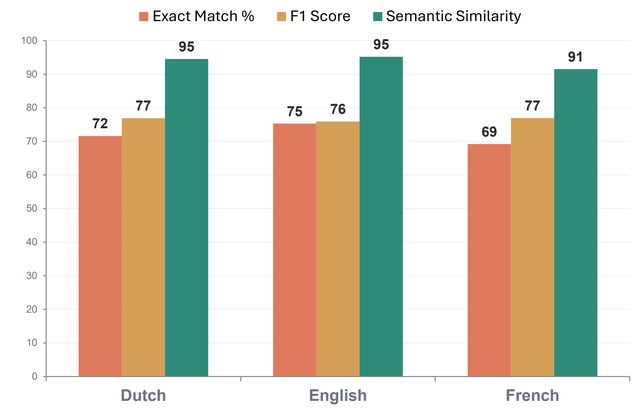


Figure 1. Comparison of exact match %, F1, and Semantic Similarity across three languages. SS consistently exceeds 91, while exact match remains below 76%.

DISTRIBUTION OF RESULTS BY LANGUAGE

Expert review reveals additional clinically valid predictions beyond exact matches.

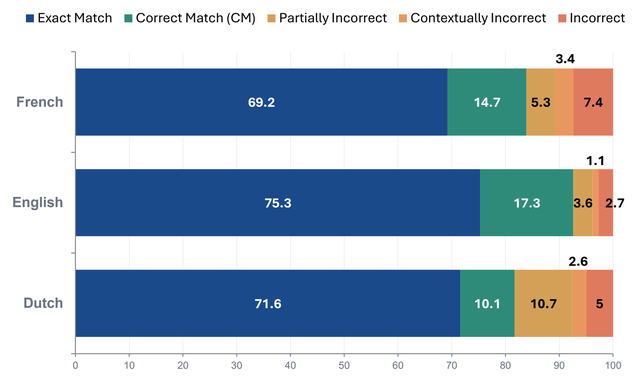


Figure 2. Human expert classification of EL predictions. CM+PIM combined represents ~81% (NL), ~96% (EN), and ~89% (FR) of all predictions — far above F1 exact match alone.

Top-k Accuracy and Mean Reciprocal Rank Across Languages

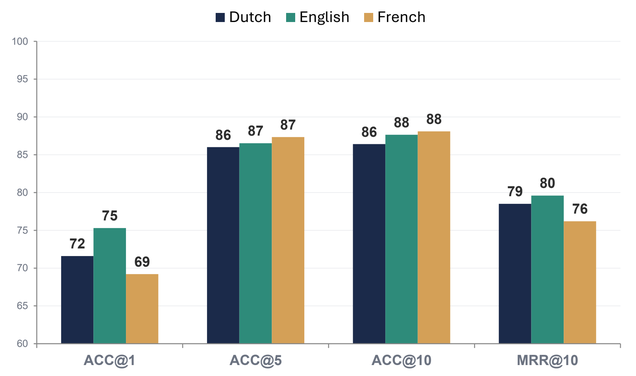


Figure 3. ACC@k shows the correct concept appears frequently in top 5–10 predictions. MRR@10 confirms it ranks highly. All metrics far exceed exact match alone.

KEY FINDINGS

1. **Exact match F1 systematically underestimates EL system performance across all three languages.**
2. On the Dutch dataset, 23.4% of predictions judged incorrect by F1 were deemed correct or partially correct by human experts.
3. Semantic Similarity scores exceeded 91 across all languages, predictions are semantically close even when CUIs differ.
4. Contextually incorrect matches (CIM) highlight the polysemy challenge in clinical abbreviations (e.g., 'PV', 'ITP', 'masse').

[1] Liu et al. (2021). Self-alignment pretraining for biomedical entity representations. NAACL-HLT.

[2] Remy et al. (2023). BioLORD-2023: Semantic textual representations fusing LLM and clinical knowledge graph insights. arXiv.

[3] Magnini et al. (2021). The E3C project: European clinical case corpus.

[4] Zanoli et al. (2024). Assessment of the E3C corpus for the recognition of disorders in clinical texts. NLE.

Contact
www.lynx.care