RWD16

An AI-powered tool to identify and assess fit for use registries for drug development and evaluation

Ghinwa Y. Hayek[1,3] ; Boris Kopin[1,3]; Sonia Zebachi[1,3]; Gaëtan Pinon[1,3]; Basile Ferry[1,3]; Elisabeth Bakker[2,3]; Alexandre Macquin[1,3]; Sieta T. de Vries[2,3]; Peter G.M. Mol[2,3]; Billy Amzal[1,3]

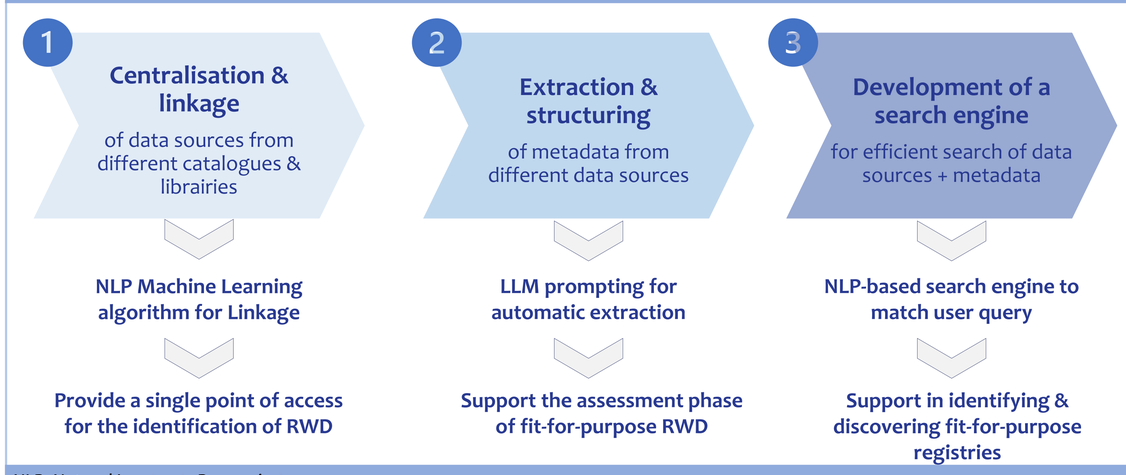
[1] Quinten Health, Paris, France; [2]University of Groningen, University Medical Center Groningen, Department of Clinical Pharmacy and Pharmacology, Groningen, Netherlands; [3] More-EUROPA Consortium

Introduction

- Selecting appropriate real-world data (RWD), particularly registries, is a primary challenge for academia, industry, regulatory, and health technology assessment (HTA) bodies, as successful submissions rely on data quality and relevance.
- The identification of a RWD is often complex due to discoverability and accessibility concerns.
- The multiplicity of data catalogues, their focus on single therapeutic areas, the fact of being behind paywalls, and the unavailability of the metadata of RWD publicly can hinder the identification of a fit-for-purpose RWD.

- **Objective: Leverage natural language processing (NLP) techniques to address these challenges by centralising and unifying patient registry data from different sources and RWD catalogues.**
- **In the context of the More-EUROPA consortium, we developed an artificial intelligence (AI) powered tool to support the identification and selection of fit-for-use real-world data across the product development lifecycle.**

Figure 1: The three key pillars of the AI-powered tool to identify and select fir-for-use RWD



NLP: Natural Language Processing
LLM: Large Language Model

Methods

- The tool was developed around **three key pillars (Figure 1)**:

➢ **Pillar 1: Centralisation, and linkages of data sources from publicly available catalogues and libraries.** Registries were extracted from:

- ✓ Data sources informed in the **HMA-EMA catalogues of RWD sources and studies.**
- ✓ Observational studies informed in **HMA-EMA catalogues of RWD sources and studies, and ClinicalTrial.gov.**
- ✓ Published papers from the **PubMed and Semantic Scholar libraries**.
- ✓ A machine learning algorithm was developed to identify, de-duplicate, and link records referring to the same registries across the different data sources (Figure 2).

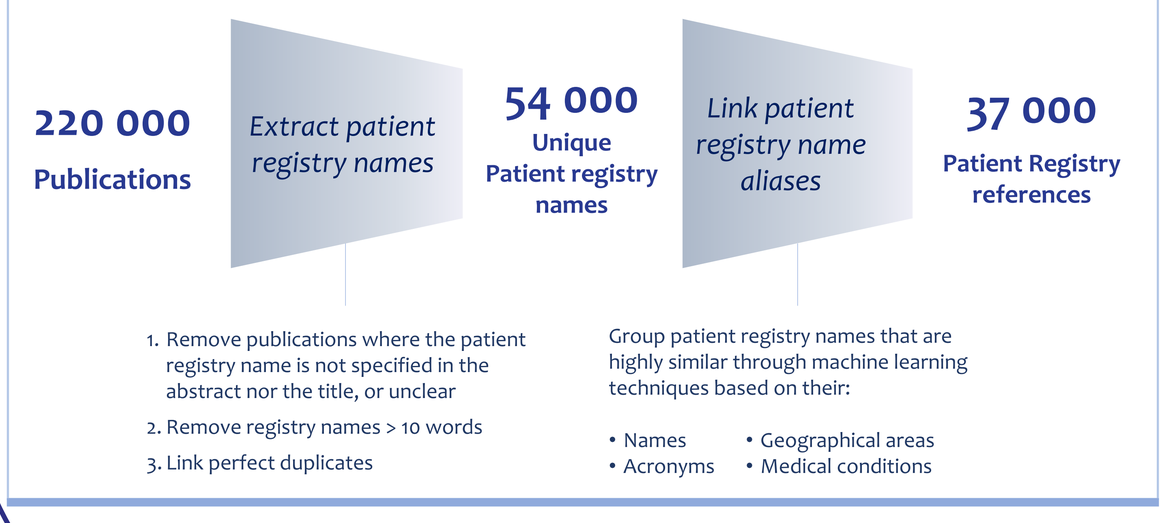
➢ **Pillar 2: Extraction of available structured and unstructured metadata across the different sources.**

- ✓ The extracted metadata was unified into a "common metadata model" based on PICOTS (population, intervention, comparator, outcome, timing, setting), leveraging machine learning and NLP techniques.
- ✓ A Large Language Model (LLM) was applied to extract key information from unstructured publications data.

➢ **Pillar 3: Development of an AI-powered search engine to support users in identifying and discovering fit-for-purpose registries within the harmonized catalogue, aligned with the user search.**

- ✓ It harnessed testing a mix of algorithms either keyword based, semantic based, or both.

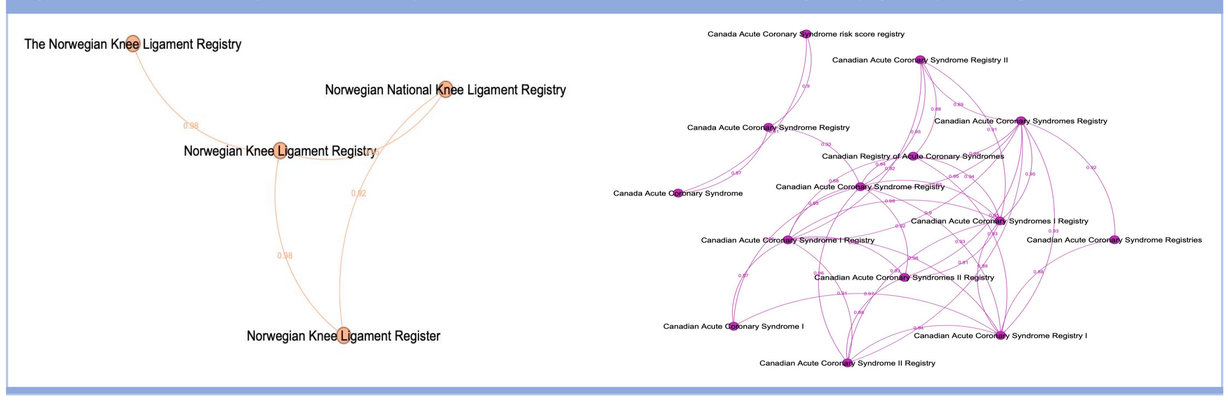
Figure 2: Extraction and linkage of registry names from publications



Results

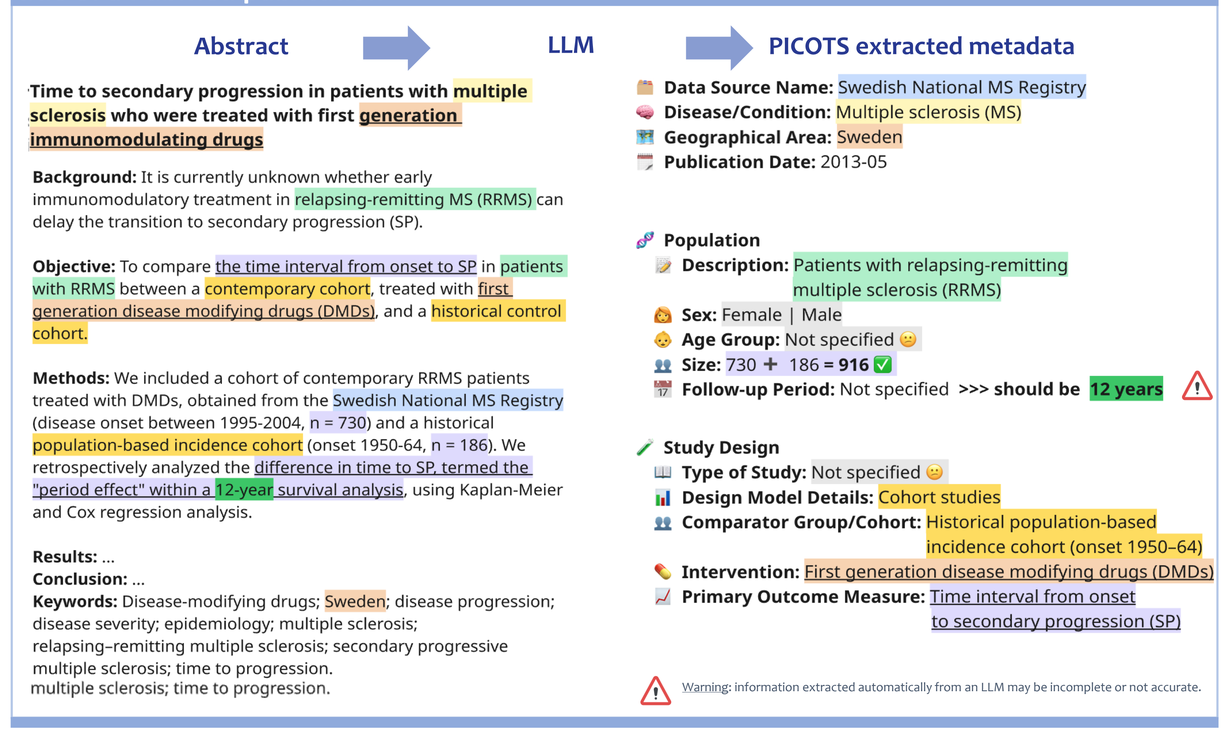
- The results of the combination of the three **pillars is an AI-powered, single-point-of-access web tool to rapidly identify, pre-assess, and pre-select registries for regulatory and HTA purposes. It results in two main deliverables constituting its backbone:**
  - ➢ A comprehensive, unified registry database consolidating metadata from multiple international data sources, powering
  - ➢ An AI-assisted interface enabling users across the product development lifecycle to search, filter, identify, and assess fit-for-use registries through an intuitive, NLP-enhanced search experience.
- **For Pillar 1: The current version includes 267 data sources extracted from the HMA-EMA catalogues of RWD sources and studies, 8,300 observational studies from Clinicaltrials.gov and HMA-EMA catalogues of RWD sources and studies, and 220,000 publications from PubMed and Semantic Scholar.**
  - ➢ The deduplication process of 220,000 publications included numerous steps (Figure 2).
  - ➢ It led to a final count of 37,000 registry references, with a 95% precision in linkages across different registries (Figure 3).

Figure 3: Two examples of multiple aliases of a same patient registry grouped together



- **For Pillar 2: The trained LLM demonstrated consistent and accurate extraction of PICOTS-related metadata from publications.**
  - ➢ Based on a manual annotation dataset of a sample of 300 publications, the accuracy for extraction for the following metadata was: **90%, 92%, 98% and 90% for registry name, medical condition, geographical area, and outcome measures,** respectively (Figure 4).
- **For Pillar 3: Multiple experiments via dataset annotations were done to evaluate our search engine, which is based on a mix of hybrid search of semantics and keywords.**
  - ➢ The optimisation of different combinations of algorithms demonstrated that a hybrid approach overperforms standalone semantic search or keyword search, leading to improvement in the quality of results following a search query.
  - ➢ The inclusion of relevant metadata (such as acronym of the RWD) in the search algorithm, improved the accuracy, thus facilitating retrieval of registries matching the search query.

Figure 4: Example of extracted metadata via LLM from unstructured data in the abstract and title of a PubMed publication



Conclusions

- The developed AI-powered tool is a comprehensive platform to identify fit-for-use registries for regulatory or HTA purposes; addressing diverse research questions across the product development lifecycle.
- Current development phases are being tested with early users, including regulators, HTA bodies, and industries, to ensure its practical adoption.

ACKNOWLEDGEMENTS: **We would like to thank all members of the More-EUROPA consortium, including the advisors, for their contributions and feedback during the different development phases of the tool. We would also like to thank Jules Perrio, Nicolas Wagner, and Salim Hassouna for their contributions to this project.**

The More-EUROPA (More Effectively Using Registries to support Patient-centered Regulatory and HTA decision-making) project has received funding from the European Union's Horizon Europe Research and Innovation Actions under grant no. 101095479 (More-EUROPA). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union nor the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

quinten health
g.hayek@quinten-health.com

MORE EUROPA
more-europa@umcg.nl
Website
LinkedIn