Automated Study Feasibility: Leveraging Agentic AI to Rapidly Identify Right Fit Secondary Data

Nicola Sawalhi-Leckenby[1], Mark Yates[1], Sophie Graham[1], Dimitra Lambrelli[1], Mireia Raluy Callado[2], Ashwin Rai[3]

[1]Thermo Fisher Scientific, London, UK; [2]Thermo Fisher Scientific, Stockholm, Sweden; [3]Thermo Fisher Scientific, Waltham, MA, USA

RWD35

Background

- Identifying appropriate data sources for real-world evidence is increasingly challenging as research questions demand granular clinical, temporal, and biomarker data. At the same time, the number and diversity of available real-world data (RWD) sources, across electronic health records, registries, claims, and linked datasets, has expanded rapidly across regions.
- The Professional Society for Health Economics and Outcomes Research (ISPOR) task force guidance emphasizes transparent, reproducible evaluation of data source suitability; however, feasibility assessments often rely on manual processes and fragmented institutional knowledge. There remains a need for scalable approaches that can operationalize ISPOR principles in a consistent and practical way across diverse data sources.
- This approach aligns with ISPOR principles of transparency, reproducibility, and fitness-for-purpose in real-world data selection.

Objectives

- This poster describes the development of a structured metadata catalogue and explores the use of an agentic artificial intelligence (AI) interface to operationalize transparent, reproducible data source selection.

Framework Development and Insights

- Data source catalogue development followed a reproducible workflow using a unified data element grid (DEG) template, harmonized from internal knowledge and EMA catalogue metadata. For initial development, 100 commonly used claims and electronic health record (EHR) databases were prioritized. DEGs were curated by subject-matter experts using a structured extraction process, with independent review by a second knowledgeable reviewer.
- This structured metadata foundation enables systematic and reproducible interrogation of data sources and forms the basis for AI-assisted querying.
- Adding an AI-assisted querying layer to curated metadata enables a more standardized, efficient, and reproducible approach to identifying fit-for-purpose data sources. The process for framework development is outlined in **Figure 1**.

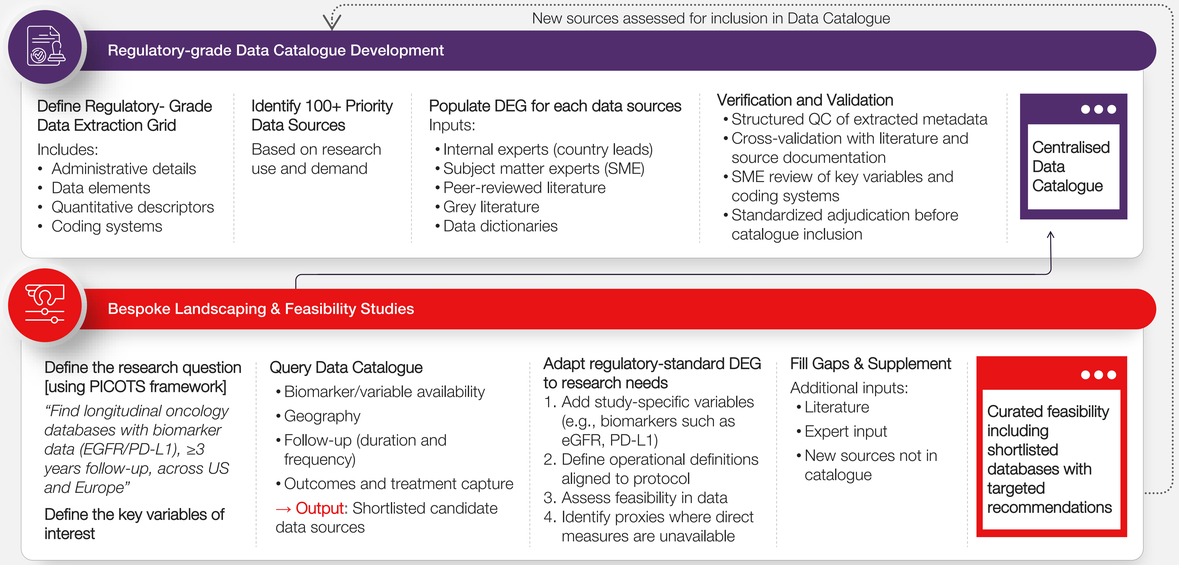
AI-Powered Data Source Recommendations

- An AI agent-based interface is under development to interpret study requirements in line with ISPOR–ISPE good practice for real-world data studies, align them with catalogue metadata, and generate standardized, transparent recommendations for data source suitability. The AI does not generate knowledge; it operationalizes expert-curated metadata.
- The planned front-end interface will allow non expert users to enter study requirements in free-text form, which are then cross checked against catalogued metadata to present data recommendations to the user. A prototype is illustrated in **Figure 2**.
- While this approach shows clear potential, development and early prototyping identified several practical challenges (**Figure 3**).
- These findings are directly informing refinement of the metadata catalogue and highlighting where additional standardization and training data are needed to improve performance.

Conclusions

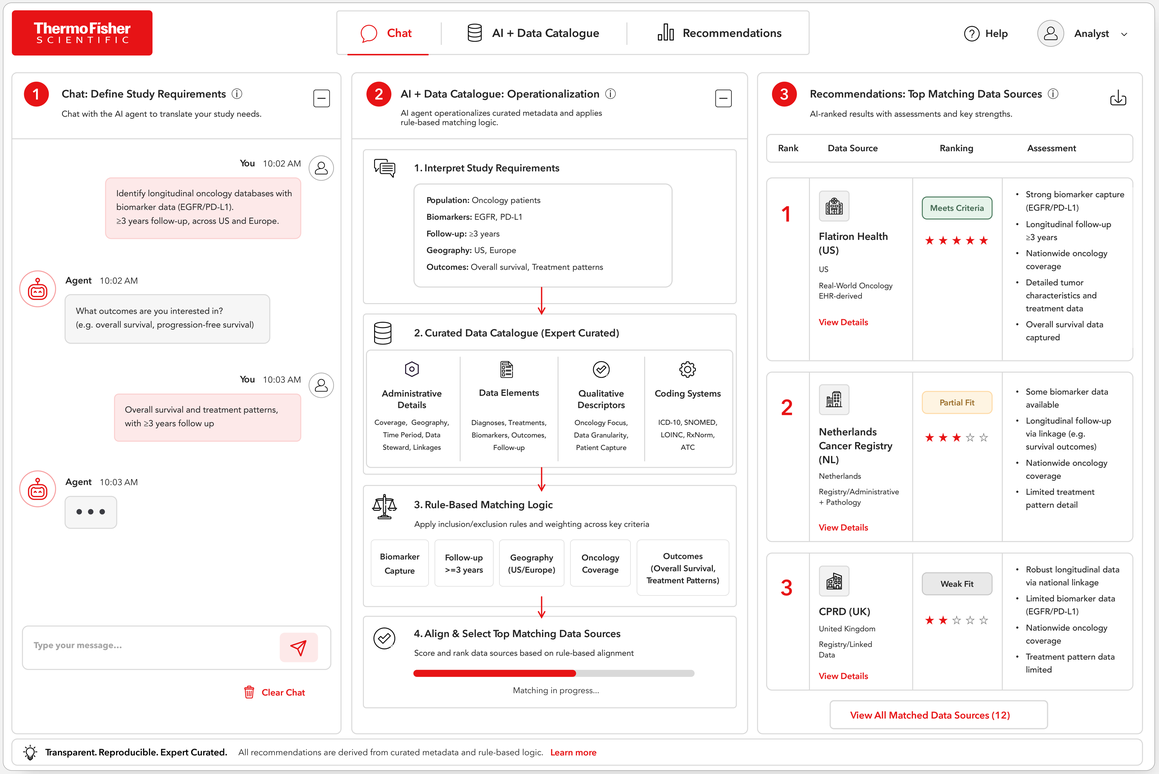
- A harmonized metadata foundation combined with an AI-driven interface provides a scalable approach to operationalizing ISPOR principles for data source selection supporting greater transparency, reproducibility, and confidence in real-world evidence studies.
- As the field continues to evolve, high-quality background data remain critical. The model will improve with further refinement and expanded data, but human expertise remains essential to guide interpretation and appropriate use.
- Ongoing work focuses on improving classification of complex data sources, expanding catalogue coverage, and evaluating the impact of AI-assisted feasibility assessments on study planning efficiency and decision quality.

Figure 1. Framework for Systematic Data Catalogue Development and Bespoke Landscaping



Abbreviations: DEG = data extraction grid; EGFR = epidermal growth factor receptor; HRD = homologous recombination deficiency; PD-L1 = programmed death ligand 1; PICOTS = population, intervention, comparator, outcome, time; rwPFS = real-world progression free survival; QC = quality control; SME = subject matter expert.

Figure 2. AI Agent–Enabled Study Recommendations



Abbreviations: AI = artificial intelligence; ATC = Anatomical Therapeutic Chemical; CRPD = Clinical Practice Research Datalink; EGFR = epidermal growth factor receptor; ICD-10 = International Classification of Diseases, 10*th* Revision; LOINC = Logical Observation Identifiers, Names, and Codes; NL= Netherlands; PD-L1 = programmed death ligand 1; SNOMED CT = Systematized Nomenclature of Medicine—Clinical Terms.

Figure 3. Key Considerations for AI-Driven Study Recommendations

Quality, standardization and validation

AI performance depends on high-quality, expert-curated metadata. Inconsistent or non-standardized data limits reliability. Validation against expert decisions and across use cases is required to ensure accuracy and consistency.

Complex study designs

Complex designs (e.g., hybrid controls, external controls, multi-cohort studies) require in-depth assessment of comparability (e.g., population, covariates, data capture) and are not amenable to simple screening approaches.

Data heterogeneity across sources

Variability in data structure, coding systems, and capture across sources (e.g., multi-country datasets, differing formats) complicates harmonization and cross-database comparisons.

Usability for non-expert users

Clear, intuitive outputs and guidance are essential for effective use by non-experts.

Automation and scalability

AI allows the possibility to automatically update the information in a regular cadence and scalability once it can access external knowledge

References

1. Berger ML, et al. *Value Health*. 2017;20(8):1003-1008.
2. Fleurence RL, et al. *Value Health*. 2024;27(6):692-701.

Disclosures

The authors are employees of PPD™ Evidera™ Real-World Data & Scientific Solutions, Thermo Fisher Scientific, and report no other relevant financial relationships.

Acknowledgments

Editorial and graphic design support were provided by Michael Grossi and Shani Berger of Thermo Fisher Scientific.