Validating a Scalable Approach to Data Fitness-for-Purpose: Database Diagnostics Applied to LEGEND-T2DM

Clair Blacketer, MPH[1,2,4], Patrick B. Ryan, PhD[1,3,4], Marc Suchard[1,5], Martijn J. Schuemie, PhD[1,5], Peter R. Rijnbeek, PhD[1,2]

1. OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA
2. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, NL
3. Department of Biomedical Informatics, Columbia University, New York, NY, USA
4. Johnson & Johnson, Raritan, NJ, USA
5. Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA, USA

Background

The success of any federated network study hinges on the first and most essential task: efficiently selecting data sources capable of informing the research question. In distributed research networks like OHDSI, where individual-level data cannot be pooled centrally, study feasibility and data source selection must rely on aggregate-level metadata(1,2). While data quality frameworks like the OHDSI Data Quality Dashboard assess conformance and plausibility, they do not evaluate whether a database contains sufficient information to inform a specific research question(3).

To address this need, we developed Database Diagnostics—a scalable, transparent, and privacy-preserving approach for evaluating whether a given data source is fit-for-purpose for a specific study. In this context, fit-for-use refers to identifying databases that contain the necessary data elements to inform the planned analysis. Importantly, the method is designed to operate on a single, standardized set of extracted summary statistics that can be reused across multiple research questions, eliminating the need for additional communication rounds during early-stage feasibility assessment. To validate the accuracy of this approach to identify potentially informative databases, we applied it retrospectively to LEGEND-T2DM, a well-established federated comparative effectiveness and safety study that evaluated the real-world performance of second-line treatments for type 2 diabetes across multiple international data sources(4).

Methods

Database Diagnostics compares a study's design requirements to precomputed summary

statistics from each database(5,6). It applies a structured library of independent rules, grouped into:

1. **Concept Coverage**: Determines whether the required concepts are present in the database and at sufficient density.
2. **Criteria Availability**: Assesses whether the database contains the clinical elements required to operationalize the study. This group is broken into two subcategories:
    - **2.1 Required criteria** are those that apply to *everyone* in an analysis and are most often used to assess the criteria that describe the target and comparator cohorts.
    - **2.2 Desired criteria** are those that apply to only *some* of the patients in an analysis and are most often used to make sure the criteria that describe the outcome are present in the database.
3. **Temporal Distribution**: Evaluates whether the database provides sufficient temporal coverage.

Validation Using LEGEND-T2DM

To evaluate the accuracy of the Database Diagnostics method to identify informative databases, we conducted a validation exercise using the publicly available results of the LEGEND-T2DM class-vs-class study. This large-scale, federated observational study systematically compared the effectiveness and safety of four second-line drug classes used to treat type 2 diabetes mellitus (T2DM): DPP4 inhibitors (DPP4i), GLP1 receptor agonists (GLP1RA), SGLT2 inhibitors (SGLT2i), and sulfonylureas (SU). The study protocol, cohort definitions, and analytic code were developed and executed using OHDSI tools, with full documentation available online [https://ohdsi-studies.github.io/LegendT2dm/Protocol].

Databases Included in the Analysis

A total of 36 databases generated and submitted a Database Profile to the OHDSI Coordinating Center in the spring of 2023 as part of a broader initiative to support simultaneous execution of four federated network studies. For the current analysis, Database Diagnostics results were available for 31 of these databases; five were excluded due to data quality issues or incomplete metadata. Among the remaining 31 databases, 12 had complete cohort diagnostics results available for both the LEGEND-T2DM exposure and outcome cohorts, allowing for full validation of the Database Diagnostics classifications against the LEGEND-T2DM reference.

Results:

Among the 12 databases with full LEGEND-T2DM cohort diagnostics available, Database Diagnostics demonstrated perfect concordance with LEGEND-T2DM in identifying drug classes with any exposure (100% agreement across all four classes). However, at the ≥1,000 persons threshold, Database Diagnostics tended to overestimate the number of databases that could construct sufficiently sized exposure cohorts, as shown in figure 1. For example, 91.7% of databases were identified by Database Diagnostics as having ≥1,000 persons exposed to GLP1RA, but only 58.3% were confirmed by LEGEND-T2DM. A similar discrepancy was seen for SGLT2I (100% vs. 66.7%).

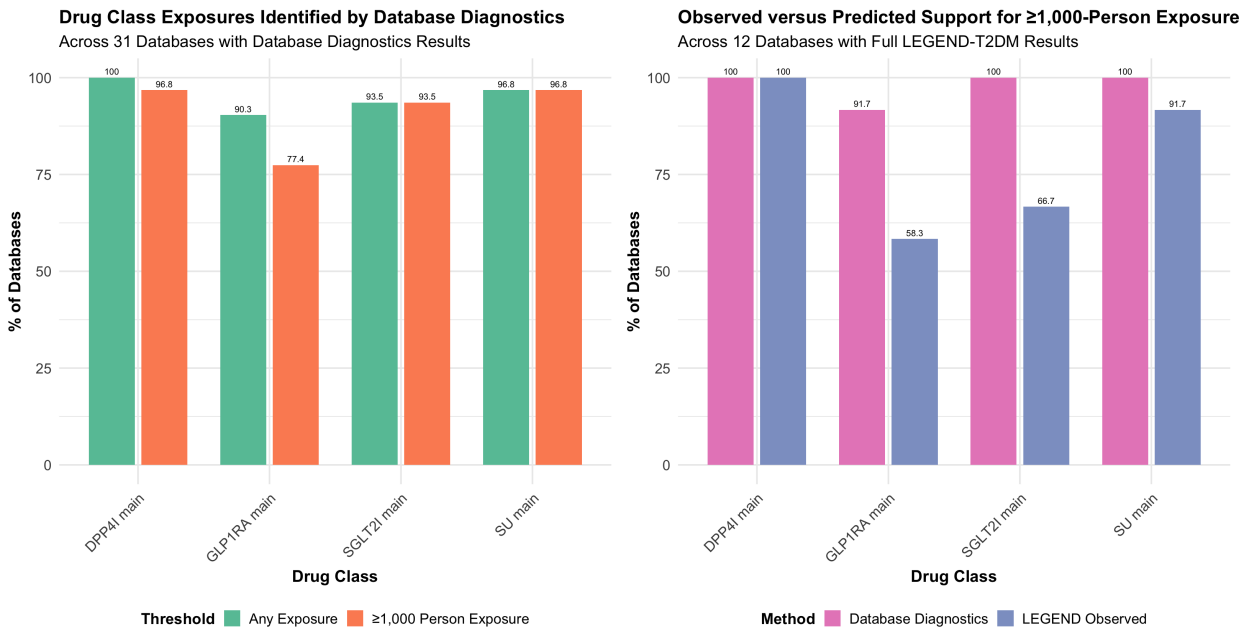

Figure 1: Comparison of drug class exposure identified by Database Diagnostics versus fully instantiated LEGEND-T2DM phenotypes. Left: 31 databases evaluated using Diagnostics. Right: 12 databases with full LEGEND-T2DM results compared at the ≥1,000-person threshold.

Database Diagnostics were evaluated against LEGEND-T2DM outcome cohort diagnostics for identifying outcome support in 12 databases. The method demonstrated high accuracy, with a sensitivity of 100.0% and a specificity of 92.0%. Positive predictive value (PPV) was 98.5%, and negative predictive value (NPV) was 100.0% (figure 2).

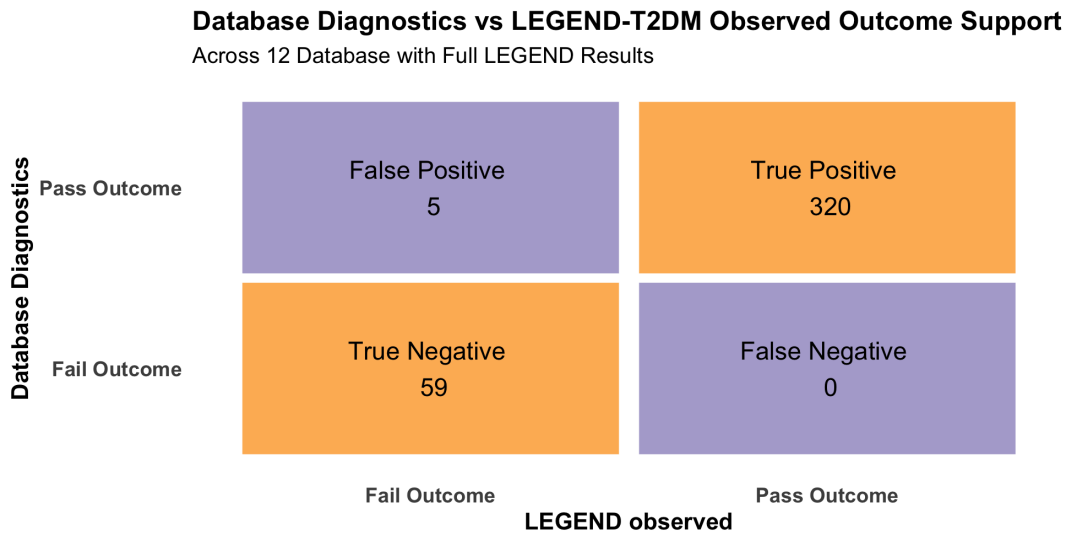

Figure 2: Confusion matrix comparing Database Diagnostics outcome predictions for 12 databases against LEGEND T2DM results.

Discussion

Database Diagnostics enables reproducible, study-specific feasibility assessments across a federated network without accessing patient-level data. Its application to LEGEND-T2DM demonstrates that it can effectively identify whether exposure and outcome cohorts are likely to be constructible in a given data source.

Discrepancies at the ≥1,000-person exposure threshold were largely due to study-specific design decisions in LEGEND-T2DM. The exposure cohort definitions required patients to have no prior exposure to other second-line antihyperglycemics, significantly narrowing cohort size compared to the broader estimates used by Database Diagnostics, which operates only on high-level concept representation and basic demographic thresholds. This highlights an important consideration: while Database Diagnostics effectively flags whether relevant concepts exist in sufficient quantity, it does not account for study-specific cohort attrition criteria.

Conclusion

Database Diagnostics offers a scalable, privacy-preserving method to assess data fitness-for-use in federated research. Its validation against LEGEND-T2DM supports its utility in accelerating study design, improving transparency, and guiding database selection. As federated networks grow in size and popularity, tools like Database Diagnostics will be essential for efficiently determining whether data sources can inform targeted research questions.

References

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.

2. Rijnbeek PR, Schuemie MJ, van der Lei J, al et. The European Health Data & Evidence Network (EHDEN): building a federated network to accelerate research. European Journal of Epidemiology. 2020;35(6):613–7.

3. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association : JAMIA. 2021 Jul 27;

4. Khera R, Schuemie MJ, Lu Y, Ostropolets A, Chen R, Hripcsak G, et al. Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies. 2022 Jun 1 [cited 2025 Jun 17]; Available from: https://bmjopen.bmj.com/content/12/6/e057977

5. Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES).

6. OMOP CDM Database Diagnostics Utility [Internet]. [cited 2023 Jun 8]. Available from: https://ohdsi.github.io/DbDiagnostics/index.html