Strengthening the Foundations of Real-World Evidence



Advancing data standardization quality, and fitness-for-purpose in federated data networks

Margaret S Blacketer

Propositions

Strengthening the Foundations of Real-World Evidence: Advancing data standardization quality, and fitness-for-purpose in federated data networks

1. The usefulness of standardized data for evidence generation depends on accurate vocabulary mapping and remains an active area of methodological development. (this thesis)
2. Seemingly minor differences in defining observable time can materially alter effect estimates, making transparency in temporal assumptions essential for valid inference. (this thesis)
3. Standardized data enables large-scale quality assessment across federated research networks using reproducible and transparent methods. (this thesis)
4. Conformance to a common data model enables the assessment of database fitness-for-use but does not guarantee it. (this thesis)
5. Database fitness-for-purpose can be systematically and reproducibly assessed using only aggregate statistics. (this thesis)
6. Perceived data loss during standardization often reflects limited transparency in transformation processes rather than true loss of meaningful information.
7. Fitness-for-purpose assessments should consider both data availability and the methodological assumptions required to generate valid evidence.
8. The ability to generate evidence in a federated network depends as much on governance and collaboration as on technical solutions.
9. There is limited empirical evidence linking data quality measures to their impact on study results in observational health research.
10. Florence Nightingale's call for standardized data to generate evidence that reduces preventable harm to patients remains relevant in the era of global real-world data networks.
11. One must imagine Sisyphus happy – Albert Camus

Margaret Blacketer

Strengthening the Foundations of Real-World Evidence:

Advancing data standardization, data quality,
and fitness-for-purpose in federated health data networks.

Versterking van de basis van real-world bewijs:

verbetering van datastandaardisatie, datakwaliteit
en geschiktheid voor het beoogde doel in gefedereerde gezondheidsdatanetwerken.

Margaret Sinclair Blacketer

Colophon

Strengthening the Foundations of Real-World Evidence: Advancing data standardization, data quality, and fitness-for-purpose in federated health data networks.

Margaret Blacketer

PhD thesis, Erasmus University Rotterdam, 2026

Cover design: Margaret Blacketer
Layout and Printed by: ProefschriftMaken | proefschriftmaken.nl
ISBN: 978-94-6534-441-6

© Margaret Blacketer. All rights reserved.

Strengthening the Foundations of Real-World Evidence:

Advancing data standardization, data quality,
and fitness-for-purpose in federated health data networks.

Versterking van de basis van real-world bewijs:

verbetering van datastandaardisatie, datakwaliteit
en geschiktheid voor het beoogde doel in gefedereerde gezondheidsdatanetwerken.

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. ir. A.J. Schuit

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Wednesday 08 July 2026 at 13:00 hrs

by

Margaret Sinclair Blacketer
born in Chesapeake, Virginia, U.S.A.

Erasmus University Rotterdam

DOCTORAL COMMITTEE:

Promotor:	Prof. dr. ir. P.R. Rijnbeek
Co-promotor:	Dr. M.J. Schuemie
Assessment Committee:	Dr. L.B. Koppert Prof. dr. R.W. Selles Prof. dr. N. Pratt

To my boys, who have walked every step of this path with me

TABLE OF CONTENTS

General Introduction 9

Data Standardization 15

Data Quality 16

Fitness for Purpose 17

Data Standardization 25

Chapter 1- European Health Data & Evidence Network – Learnings from Building out a Standardized International Health Data Network 27

Chapter 2- Evaluation of the Impact of Defining Observable Time in Real-World Data on Outcome Incidence 51

Data Quality 73

Chapter 3 - Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality 75

Chapter 4 - Using the Data Quality Dashboard to Improve the EHDEN Network 93

Fitness-for-Purpose 113

Chapter 5 - Building the Largest Federated Health Data Network in Europe: Advancing Collaboration and Real-World Evidence Generation 115

Chapter 6 - Validation of the Database Diagnostics method: Assessing Fitness-for-Purpose in Federated Networks 152

General Discussion 181

Summary 193

Samenvatting 196

Acknowledgements 200

Curriculum Vitae 203

PhD Portfolio 204

List of Publications 205

About the Author 212

General Introduction

Clair Blacketer

A philosophical debate rages in the medical literature. Two eminent scientists publish mortality rates compared across multiple hospitals using a novel statistical method. Mathematicians decry the potentially misleading nature of the reporting while physicians denounce the radical conclusions of the article suggesting hospitals should be completely redesigned. The authors respond, defending their choices in a point-by-point rebuttal while at the same time disparaging the quality and collection of the data utilized. A decade and a half before the advent of the light bulb, informaticists had already pinpointed one of the key challenges in using observational health data to generate evidence.

In 1863 Florence Nightingale published the third edition of "Notes On Hospitals", a pioneering epidemiological text that utilized hospital mortality statistics provided by the Registrar General, William Farr, to support her claim that urban hospitals were unsanitary [1–3]. As reported, London hospitals had a 90.84% mortality rate, a ghastly number that rightly created an uproar among the medical community at the time [4–7]. Farr nobly defended both Miss Nightingale and his methods in the press, revealing that the true misunderstanding came from a lack of clarity in the presentation of the results (deaths per person year rather than percent of total patients) [8,9]. In the course of his defense, Farr criticized the quality of the hospital data, saying "… information in the London hospitals was exceedingly defective at that time, and I suggested that competent hospital registrars should be appointed" [8]. He understood that even the best statistical approach can produce questionable results if the underlying data are not handled appropriately.

Concerns over data quality have long plagued the field of medical informatics, especially since observational data like that Farr and Nightingale used are not originally collected for the purposes of research [10,11]. Nightingale herself lamented "…accurate hospital [data] are much more rare than is generally imagined…" [1]. Today, the field has evolved from relying on data from hospitals in London to utilizing data from millions of patients around the world. However, more than 150 years later, the goals of the nineteenth century to produce trustworthy, actionable information to improve patient care remain only partially realized. While experts are still working towards the generation of robust, reliable, reproducible, repeatable, and generalizable, evidence, this ambition rests on the fundamental truth that it all begins with the data. Achieving this requires more than simply collecting information. We must ensure standardized structure and content, evaluate and improve data quality, and assess data fitness-for-purpose such that evidence can be generated confidently in a way that inspires trust from clinicians, researchers, and the patients they serve.

AIMS OF THIS THESIS

With these goals in mind, this thesis focuses on improving the foundational elements required to generate trustworthy evidence from observational health data at scale. Specifically, it examines how to advance the science of data standardization, data quality assessment, and evaluation of database fitness-for-purpose within federated observational health data networks. As key underpinnings of evidence generation, these areas are deeply connected. Standardizing the structure and content of data supports consistent analytic methods across disparate sources. Standardization is also critical to the development and application of broad data quality assessments in federated networks to ensure that analyses are based on complete and accurate information. When looking for data sources to answer specific study questions, fitness-for-purpose evaluations must leverage these standardized, high-quality data sources in ways that do not jeopardize patient privacy. Exploring these ideas, each chapter of this thesis addresses a critical step toward building a more trustworthy, reliable, and scalable evidence generation process.

OBSERVATIONAL HEALTH DATA

Over the past few decades, digitization in healthcare has led to a massive increase in the amount of observational health data available for research [12–14]. Collectively known as real-world data (RWD), electronic health records (EHRs), insurance claims, and patient registry information are generated from routine clinical care and administrative processes. Unlike data from randomized controlled trials (RCTs) which are collected in experimental settings with research as their primary purpose, these data are recorded as patients are treated, making them a byproduct of day-to-day practice. The depth and breadth of observational health data are unmatched, covering diverse populations, long spans of time, and a wide array of health outcomes. Health questions that would be impractical or unethical in randomized trials are now possible for researchers to study [15,16].

Observational health data are increasingly considered important resources for generating real-world evidence (RWE) to inform clinical, regulatory, and policy decisions [17–19]. Studies of treatment effectiveness, safety surveillance, and disease epidemiology, among others, are faster and the results complementary to clinical trials as they use secondary data from settings that reflect actual patient care [20–22]. When evidence from trials is insufficient, professional medical societies are starting to incorporate high-quality observational studies into clinical practice guidelines [23]. Likewise, regulatory agencies such as the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) now routinely consider real-world evidence in drug approval and pharmacovigilance efforts [24–28]. Pharmaceutical companies, medical device manufacturers, and health technology firms utilize observational data to guide product development, assess market access,

and meet post-authorization study requirements [25,26]. Researchers across the globe have found success in using these data to innovate on and refine analytic approaches, methodology, and study designs. This, in turn, has led to steady improvement in the validity, reproducibility, and (critically) interpretability of evidence [29–31]. Observational data can and will continue to broaden both the scope and applicability of health outcomes research at a variety of levels throughout the larger healthcare ecosystem.

At the same time, the inherent nature of observational data can present challenges. They are often heterogeneous in form as well as patient representation due to the fact these data are collected for clinical or administrative purposes rather than for research [32,33]. Different providers, institutions, and countries use varying terminologies, data formats, and recording practices, making interoperability, reproducibility and consistent interpretation difficult [34,35]. Furthermore, individual databases often only capture a part of a patient's total healthcare journey i.e., encounters within a single hospital system or claims from one insurer, leading to incomplete records and the potential for bias. Addressing these gaps, researchers need the opportunity to study different populations and ensure the consistency of findings across care settings by including multiple data sources in analysis. Large-scale data networks have emerged as a way to fill this need, bringing disparate databases together for collaborative research [36–39]. Due to their ability to preserve patient privacy while at the same time maximizing the scientific value of RWD, the federated approach to building observational health data networks has become increasingly popular. In a federated network, data are transformed to a standard format while remaining securely held behind institutional firewalls. This practice allows the same analytic code to be executed across all sites without sharing patient-level information, protecting privacy yet still allowing researchers to learn from a variety of sources. Federated networks create an incredibly rich data environment that is substantially more complete and nuanced than any single existing data source could provide.

FEDERATED DATA NETWORKS

Federated networks represent one approach to bringing together real-world data, but they are part of a broader set of data network designs. In general, networks are organized to connect disparate databases so that evidence can be generated across populations, care settings, and health systems. There are several ways networks can be structured: centralized, federated, and hybrid. Centralized models pool patient-level data into a single repository while federated models have data remain under local control with only aggregated study results shared [40,41]. Hybrid models retain some elements of centralization combined with distributed analytics [41–44]. Each design requires trade-offs between efficiency, governance, and privacy, but all are motivated by the belief that

scientific knowledge is reinforced when evidence is drawn from multiple, heterogenous data sources.

Lessons from Early Initiatives

Several early, large-scale federated initiatives have demonstrated the feasibility of this approach while also pinpointing inherent challenges. In the United States, the FDA Sentinel System was among the first to show how distributed data and analytics could support continuous post marketing safety surveillance. Sentinel established a network of healthcare data partners who retained their patient level data but converted it into a common model, enabling standardized queries to be run across sites. This allowed regulators to monitor the safety of medical products in near real time without requiring centralization of sensitive data [45,46].

In Europe, projects such as the European Union - Adverse Drug Reactions (EU-ADR), Innovative Medicines Initiative – European Medical Information Framework (IMI-EMIF) and Innovative Medicines Initiative - Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium (IMI-PROTECT) explored similar questions of how to use existing observational data sources to monitor drug safety across countries [37,38,47,48]. These efforts showcased the potential of combining data to support drug safety monitoring but also reinforced the requirement that data must be transformed into a standardized format for analyses to be both efficient and scalable. To be successful, both the structure (how data are organized) and the content (the representation of clinical events, diagnosis, and medications) need harmonization.

Toward a Federated Future

Building on these experiences, the European Health Data and Evidence Network (EHDEN) was initiated to establish a fully federated research system across Europe [49,50]. EHDEN connects over 200 data sources from 29 countries, each retaining their data locally, adhering to institutional governance requirements (figure 1). The network enables large-scale, privacy-preserving evidence generation across hundreds of millions of patient records through distributed analytics that rely on standardized data. EHDEN represents a new generation of data networks, one that respects governance constraints while ensuring that data can be used to maximum patient benefit.

DATA STANDARDIZATION

EHDEN demonstrates that the power of federated networks lies not only in connecting data partners through a common goal but ensuring they can be analyzed consistently through data standardization. Without a shared structure and vocabulary, analyses conducted across different data sources would remain fragmented and incomparable. Common data models (CDMs) provide this foundation by defining how data are organized and how clinical ideas are represented (figure 2).

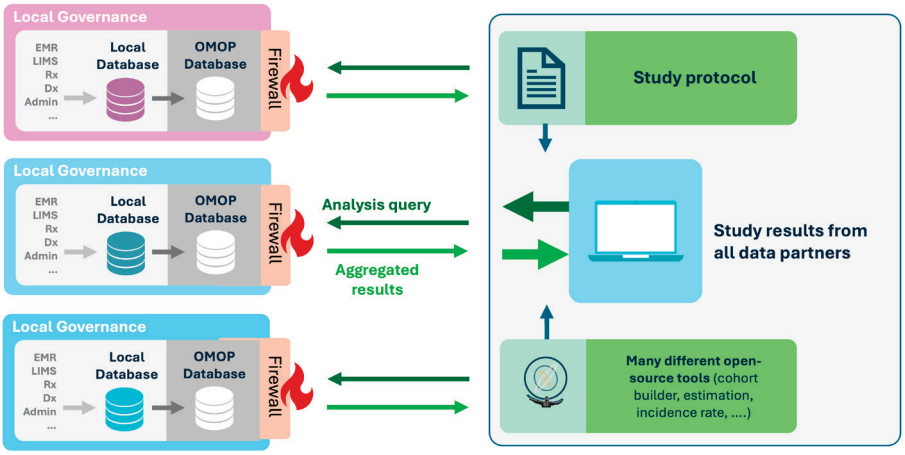


Figure 1: Overview of a federated data network, how queries are shared and results aggregated.

Among available CDMs the Observational Medical Outcomes Partnership (OMOP) CDM has emerged as a worldwide leading standard for secondary research [51,52]. Unlike other models designed primarily for data exchange and clinical care, such as Fast Healthcare Interoperability Resources (FHIR), the OMOP CDM is purpose built for observational research at scale [53]. By mapping data into OMOP, researchers gain the ability to apply analytic methods uniformly across harmonized data sets, which facilitates not only efficiency but also methodological innovation. Large-scale studies conducted by the Observational Health Data Sciences and Informatics (OHDSI) community such as Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM) and LEGEND Hypertension have shown how OMOP standardized data can support rigorous comparative effectiveness and safety research [54,55]. Similarly, the OHDSI Health Analytics Data to Evidence Suite (HADES) of open-source methods relies on the OMOP CDM to support advances in the areas of characterization, causal inference, and predictive modeling, among others [56].

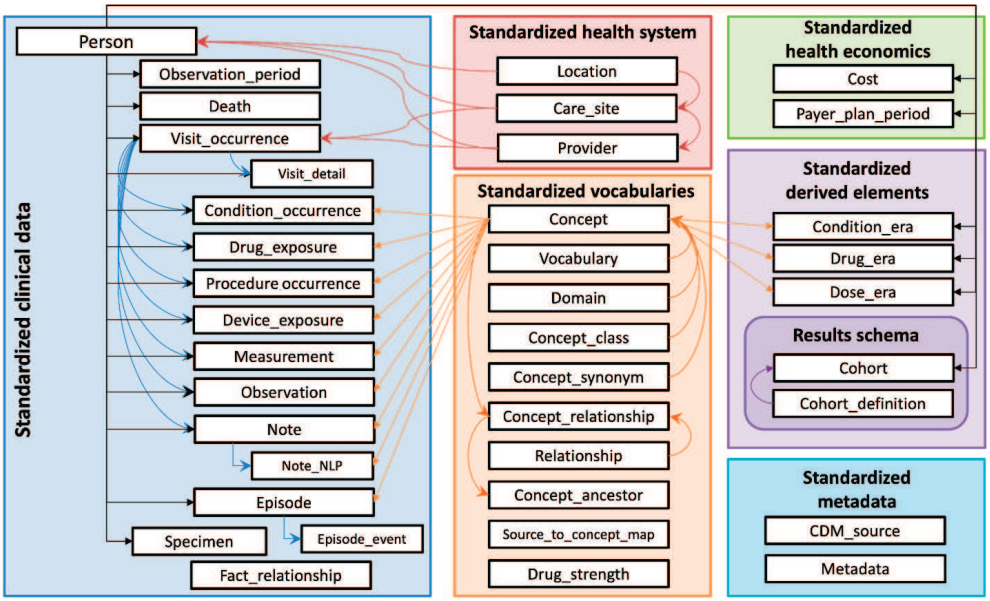


Figure 2: The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) v5.4.

This thesis improves the science of data standardization, investigating not only how data should be harmonized but also how those processes influence the reliability of the evidence produced. **Chapter 1** focuses on EHDEN and the extract–transform–load (ETL) process to map data to the OMOP CDM. This study identifies the challenges encountered during standardization of sources in the network and the lessons learned from implementing ETL pipelines across diverse countries, healthcare systems, and technical environments. **Chapter 2** then examines a methodological issue that comes up both during the ETL process and in the design of epidemiologic studies. How to define a patient's observable time, or the time during which a patient's records are reliably captured, is a decision that has potentially broad implications for study outcomes. Different approaches can lead to bias in incidence rate estimates, which in turn raises questions about how data standardization may uncover such issues while also providing opportunities to address them.

DATA QUALITY

Even after data have been standardized into a common model there is still no guarantee that they are of sufficiently high quality to support robust evidence generation [57–60]. The concept of data quality itself is elusive. Despite decades of discussion, there is no single definition universally accepted in the field. An early suggestion from Juran defines data quality as "data that are fit for use in their intended operational, decision making,

planning, and strategic roles" [61]. More recently, Patricia Brennan, former director of the U.S. National Library of Medicine, described observational health data quality as the "… problem of ensuring the validity of the clinical record as a representation of the true state of the patient" [62]. This thesis will explore these perspectives separately as they highlight different interpretations of the term data quality. "Fitness-for-purpose" is the extent to which a dataset contains the required elements to answer a particular scientific question, while "Data quality" is the extent to which those elements themselves are reliable. This thesis explores both, examining data quality first in chapters 3 and 4 and then data fitness-for-purpose in chapters 5 and 6.

Understanding and assessing data quality is especially important in federated data networks. There must be ways to remotely measure and evaluate quality in a standardized and transparent manner that does not jeopardize patient privacy as conformance, completeness, or plausibility errors in a data source can introduce bias in results [62–64]. Transparent and consistent quality assessment allows researchers to identify limitations and adjust study designs as needed to reduce this potential for bias and improve the reliability of the evidence generated.

This thesis explores these ideas through two studies. **Chapter 3** describes the development of the Data Quality Dashboard (DQD), an open-source tool designed to systematically evaluate OMOP CDM standardized databases against a comprehensive set of quality checks. This chapter describes the principles and processes behind the DQD, its implementation, and how the results can be used to identify potential quality issues. **Chapter 4** applies the DQD to the databases in EHDEN, illustrating how quality assessments can be conducted at scale across a federated network to improve the network as a whole.

FITNESS FOR PURPOSE

In the previous section, data quality was defined as the extent to which data faithfully represent the true state of the patient. Fitness-for-purpose complements this idea by determining whether a database contains the specific elements required to answer a particular scientific question. A data set may be of generally high quality, yet if it does not reliably capture the exposures, outcomes, or follow up time needed, it cannot be considered suitable for that analysis. Fitness-for-purpose is therefore not an inherent property of a database but a study-specific assessment.

Existing approaches to evaluating fitness-for-purpose have relied on study-specific diagnostics, such as OHDSI's Cohort Diagnostics and Study Diagnostics tools [65,66]. These tools enable detailed inspection of both phenotype definitions and demographic distributions, but they require study code to be executed across each participating data-

base. In large, federated networks, this process can be resource intensive and inefficient, particularly when hundreds of potential databases must be considered (as is the case in EHDEN). The lack of scalable methods to rapidly assess whether a study can reasonably be run in a given database represents a critical barrier to reproducible research at scale.

In this thesis**, Chapter 5** describes the databases in EHDEN, highlighting the scale, diversity, and the metadata the network provides, collected from each data source. **Chapter 6** then introduces the Database Diagnostics method, showing how that metadata can be used to support fitness-for-purpose assessments without needing access to patient-level data. The method is validated using the LEGEND-T2DM study as a reference case, demonstrating that Database Diagnostics offers a scalable, reproducible, and privacy preserving approach for evaluating whether databases can support specific research questions.

SCOPE OF THIS THESIS

This thesis is organized into 6 chapters. **Chapter 1** examines the ETL process utilized within EHDEN, documenting the challenges and lessons learned from large-scale standardization across heterogenous databases representing different countries and data environments. **Chapter 2** highlights the methodological implications of different approaches to defining observable time within standardized data sources, showing how they affect study outcomes. Chapters 3 and 4 focus on data quality; **chapter 3** describes the development of the DQD as a systematic tool for evaluating the quality of OMOP CDM standardized databases, while **chapter 4** applies the DQD across the EHDEN network to illustrate the feasibility and value of coordinated quality assessment. Finally, chapters 5 and 6 address fitness for purpose. **Chapter 5** provides an overview of the EHDEN network itself, including its scale and diversity, and **chapter 6** introduces and validates the database diagnostics method as a viable approach for assessing study specific database fitness-for-purpose in federated networks like EHDEN. Together, these studies advanced the science of data standardization, data quality, and fitness for purpose in federated health data networks, laying the groundwork for more reliable and efficient real world evidence generation.

REFERENCES

1 Nightingale F. *Notes on Hospitals*. 3rd ed. London: Longman, Green, Longman, Roberts, and Green 1863.

2 Hill GB. "On prognosis" by William Farr (British Medical Almanack 1838; Supplement 199-216). Part 2 (pages 208-216). *Soz Praventivmed*. 2003;48:279–84. doi: 10.1007/s00038-003-3005-7

3 Farr W. "On Prognosis" by William Farr (British Medical Almanack 1838; Supplement 199–216). *Sozial- und Präventivmedizin/Social and Preventive Medicine*. 2003;48:219–24. doi: 10.1007/s00038-003-3004-8

4 Holmes T. MORTALITY IN HOSPITALS. *The Lancet*. 1864;83:365–6. doi: 10.1016/S0140-6736(02)59168-6

5 Response to Letter by William Farr. *Medical Times and Gazette*. 1864;187–8.

6 Bristowe, JS. Hospital Mortality. *Medical Times and Gazette*. 1864;491–2.

7 Review: Notes on Hospitals, by Florence Nightingale. *Medical Times and Gazette*. 1864;129–30.

8 Farr W. MORTALITY IN HOSPITALS. *The Lancet*. 1864;83:420–2. doi: 10.1016/S0140-6736(02)40958-0

9 Farr W. Miss Nightingale's "Notes on Hospitals." *Medical Times and Gazette*. 1864;186–7.

10 Green J, Wintfeld N. Report Cards on Cardiac Surgeons — Assessing New York State's Approach. *New England Journal of Medicine*. 1995;332:1229–33. doi: 10.1056/NEJM199505043321812

11 Iezzoni LI. 100 apples divided by 15 red herrings: a cautionary tale from the mid-19th century on comparing hospital mortality rates. *Ann Intern Med*. 1996;124:1079–85. doi: 10.7326/0003-4819-124-12-199606150-00009

12 Mitchell JB, Bubolz T, Paul JE, *et al.* Using Medicare claims for outcomes research. *Med Care*. 1994;32:JS38-51.

13 Lewis NJ, Patwell JT, Briesacher BA. The role of insurance claims databases in drug therapy outcomes research. *Pharmacoeconomics*. 1993;4:323–30. doi: 10.2165/00019053-199304050-00003

14 Adler-Milstein J, DesRoches CM, Kralovec P, *et al.* Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. *Health Aff (Millwood)*. 2015;34:2174–80. doi: 10.1377/hlthaff.2015.0992

15 Scherer DG, Annett RD, Brody JL. Ethical issues in adolescent and parent informed consent for pediatric asthma research participation. *J Asthma*. 2007;44:489–96. doi: 10.1080/02770900701247137

16 Dash S, Shakyawar SK, Sharma M, *et al.* Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*. 2019;6:54. doi: 10.1186/s40537-019-0217-0

17 The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology. https://encepp.europa.eu/encepp-toolkit/methodological-guide_en (accessed 6 September 2024)

18 U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018.

19 Heads of Medicines Agency, European Medicines Agency. HMA-EMA Joint Big Data Taskforce Summary Report. 2019;48.

20 Williams RD, Reps JM, Rijnbeek PR, *et al.* 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. *Knee Surgery, Sports Traumatology, Arthroscopy*. Published Online First: 6 December 2021. doi: 10.1007/s00167-021-06799-y

21 Voss EA, Shoaibi A, Yin Hui Lai L, *et al.* Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 da-

tabases: a multinational retrospective cohort study. *EClinicalMedicine*. 2023;58:101932. doi: 10.1016/j.eclinm.2023.101932

22 Khera R, Aminorroaya A, Dhingra LS, *et al.* Comparative Effectiveness of Second-Line Antihyperglycemic Agents for Cardiovascular Outcomes: A Multinational, Federated Analysis of LEGEND-T2DM. *Journal of the American College of Cardiology*. 2024;84:904–17. doi: 10.1016/j.jacc.2024.05.069

23 Gershon AS, Lindenauer PK, Wilson KC, *et al.* Informing Healthcare Decisions with Observational Research Assessing Causal Effect. An Official American Thoracic Society Research Statement. *Am J Respir Crit Care Med*. 2021;203:14–23. doi: 10.1164/rccm.202010-3943ST

24 Jonker C, Bakker E, Kurz X, *et al.* Contribution of patient registries to regulatory decision making on rare diseases medicinal products in Europe. *Frontiers in Pharmacology*. Published Online First: 2022. doi: 10.3389/FPHAR.2022.924648

25 Eskola SM, Leufkens H, Bate A, *et al.* The Role of Real-World Data and Evidence in Oncology Medicines Approved in EU in 2018-2019. *Journal of Cancer Policy*. Published Online First: 2023. doi: 10.1016/J.JCPO.2023.100424

26 Bolislis WR, Fay M, Kühler TC. Use of Real-world Data for New Drug Applications and Line Extensions. *Clinical Therapeutics*. Published Online First: 2020. doi: 10.1016/J.CLINTHERA.2020.03.006

27 Bakker E, Plueschke K, Jonker C, *et al.* Contribution of Real-World Evidence in European Medicines Agency's Regulatory Decision Making. *Clinical pharmacology and therapy*. Published Online First: 2022. doi: 10.1002/CPT.2766

28 Flynn RWV, Plueschke K, Quinten C, *et al.* Marketing Authorization Applications Made to the European Medicines Agency in 2018–2019: What was the Contribution of Real-World Evidence? Published Online First: 2021. doi: 10.1002/CPT.2461

29 Ryan PB, Schuemie MJ, Gruber S, *et al.* Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36 Suppl 1:S59-72. doi: 10.1007/s40264-013-0099-6

30 Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A*. 2016;113:7329–36. doi: 10.1073/pnas.1510502113

31 Suchard MA, Simpson SE, Zorych I, *et al.* Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans Model Comput Simul*. 2013;23. doi: 10.1145/2414416.2414791

32 Trifirò G, Coloma P, Rijnbeek P, *et al.* Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *Journal of internal medicine*. Published Online First: 2014. doi: 10.1111/JOIM.12159

33 Stang PE, Ryan PB, Racoosin JA, *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010;153:600–6. doi: 10.7326/0003-4819-153-9-201011020-00010

34 Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. *Clinical Pharmacology & Therapeutics*. 2019;106:36–9. doi: 10.1002/cpt.1426

35 Gini R, Pajouheshnia R, Gardarsdottir H, *et al.* Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review. *Pharmacoepidemiology and Drug Safety*. Published Online First: 2024. doi: 10.1002/PDS.5787

36 Gini R, Schuemie M, Brown J, *et al.* Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. *eGEMs*. Published Online First: 2016.

37 Coloma PM, Schuemie MJ, Trifirò G, *et al.* Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiology and Drug Safety*. 2011;20:1–11. doi: 10.1002/pds.2053
38 Reynolds RF, Kurz X, de Groot MCH, *et al.* The IMI PROTECT project: purpose, organizational structure, and procedures. *Pharmacoepidemiol Drug Saf*. 2016;25 Suppl 1:5–10. doi: 10.1002/pds.3933
39 EFPIA Recommendations on a Connected Data System in Europe. https://www.efpia.eu/news-events/the-efpia-view/efpia-news/efpia-recommendations-on-a-connected-data-system-in-europe/ (accessed 6 September 2024)
40 Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28:427–43. doi: 10.1093/jamia/ocaa196
41 van Baalen V, Didden E-M, Rosenberg D, *et al.* Increase transparency and reproducibility of real-world evidence in rare diseases through disease-specific Federated Data Networks. *Pharmacoepidemiol Drug Saf*. 2024;33:e5778. doi: 10.1002/pds.5778
42 Fleurence RL, Curtis LH, Califf RM, *et al.* Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*. 2014;21:578–82. doi: 10.1136/amiajnl-2014-002747
43 Bardenheuer K, Van Speybroeck M, Hague C, *et al.* Haematology Outcomes Network in Europe (HONEUR)-A collaborative, interdisciplinary platform to harness the potential of real-world data in hematology. *Eur J Haematol*. 2022;109:138–45. doi: 10.1111/ejh.13780
44 Oberprieler NG, Pladevall-Vila M, Johannes C, *et al.* FOUNTAIN: a modular research platform for integrated real-world evidence generation. *BMC Med Res Methodol*. 2024;24:224. doi: 10.1186/s12874-024-02344-w
45 Behrman RE, Benner JS, Brown JS, *et al.* Developing the Sentinel System--a national resource for evidence development. *N Engl J Med*. 2011;364:498–9. doi: 10.1056/NEJMp1014427
46 Ball R, Robb M, Anderson SA, *et al.* The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99:265–8. doi: 10.1002/cpt.320
47 Lovestone S. The European medical information framework: A novel ecosystem for sharing healthcare data across Europe. *Learn Health Syst*. 2019;4:e10214. doi: 10.1002/lrh2.10214
48 Klungel O, Kurz X, Groot MC, *et al.* Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. *Pharmacoepidemiology and drug safety*. Published Online First: 2016. doi: 10.1002/PDS.3968
49 Rijnbeek PR, Schuemie MJ, van der Lei J, *et al.* The European Health Data & Evidence Network (EHDEN): building a federated network to accelerate research. *European Journal of Epidemiology*. 2020;35:613–7. doi: 10.1007/s10654-020-00606-5
50 European Health Data Evidence Network. ehden.eu. https://www.ehden.eu/ (accessed 12 March 2021)
51 Observational Health Data Sciences and Informatics. OMOP Common Data Model (CDM). https://ohdsi.github.io/CommonDataModel/
52 Voss EA, Makadia R, Matcho A, *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. 2015;22:553–64. doi: 10.1093/jamia/ocu023
53 Overview - FHIR v5.0.0. https://www.hl7.org/fhir/overview.html (accessed 3 September 2025)
54 Khera R, Schuemie MJ, Lu Y, *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. Published Online First: 1 June 2022. doi: 10.1136/bmjopen-2021-057977

55 Suchard MA, Schuemie MJ, Krumholz HM, *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet*. 2019;394:1816–26. doi: 10.1016/S0140-6736(19)32317-7
56 Schuemie M, Reps J, Black A, *et al.* Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research. *Stud Health Technol Inform*. 2024;310:966–70. doi: 10.3233/SHTI231108
57 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translat Bioinforma*. 2010;2010:1–5.
58 Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf Med*. 2009;48:38–44.
59 Otero Varela L, Wiebe N, Niven DJ, *et al.* Evaluation of interventions to improve electronic health record documentation within the inpatient setting: a protocol for a systematic review. *Syst Rev*. 2019;8:54. doi: 10.1186/s13643-019-0971-2
60 Alla F, Rosilio M, Funck-Brentano C, *et al.* How can the quality of medical data in pharmacovigilance, pharmacoepidemiology and clinical studies be guaranteed? *Therapie*. 2013;68:209–23. doi: 10.2515/therapie/2013040
61 Juran J, Gryna F, Bingham R. *Quality Control Handbook*. New York, NY: McGraw Hill 1974.
62 Brennan PF, Stead WW. Assessing Data Quality. *J Am Med Inform Assoc*. 2000;7:106–7.
63 Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*. 2013;46:830–6. doi: 10.1016/j.jbi.2013.06.010
64 Kahn MG, Brown JS, Chun AT, *et al.* Transparent reporting of data quality in distributed data networks. *EGEMS (Washington, DC)*. 2015;3:1052. doi: 10.13063/2327-9214.1052
65 Rao G, Shoaibi A, Makadia R, *et al.* CohortDiagnostics: phenotype evaluation across a network of observational data sources using population-level characterization. *medRxiv*. Published Online First: 2023. doi: 10.1101/2023.06.28.23291982
66 Conover MM, Ryan PB, Chen Y, *et al.* Objective study validity diagnostics: a framework requiring pre-specified, empirical verification to increase trust in the reliability of real-world evidence. *Journal of the American Medical Informatics Association*. 2025;32:518–25. doi: 10.1093/jamia/ocad030

Data Standardization

1

European Health Data & Evidence Network – Learnings from Building out a Standardized International Health Data Network

Erica A. Voss, MPH[1,2,3]
Clair Blacketer, MPH[1,2,3]
Sebastiaan van Sandijk, MSc[1,4]
Maxim Moinat, MSc[1,2]
Michael Kallfelz, MD[1,4]
Michel van Speybroeck, MSc[3]
Daniel Prieto-Alhambra, PhD[1,2,5]
Martijn J. Schuemie, PhD[1,3,6]
Peter R. Rijnbeek, PhD[1,2]

[1]OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States
[2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands
[3]Janssen Pharmaceutical Research and Development LLC, Raritan, NJ 08869, United States
[4]Odysseus Data Services, Prague, Czech Republic
[5]Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, United Kingdom
[6]Department of Biostatistics, University of California, Los Angeles, CA 90095, United States

Voss, E. A., Blacketer, C., van Sandijk, S., Moinat, M., Kallfelz, M., van Speybroeck, M., Prieto-Alhambra, D., Schuemie, M. J., & Rijnbeek, P. R. (2023). European Health Data & Evidence Network-learnings from building out a standardized international health data network. Journal of the American Medical Informatics Association : JAMIA, 31(1), 209–219. https://doi.org/10.1093/jamia/ocad214

ABSTRACT

Objective: Health data standardized to a common data model (CDM) simplifies and facilitates research. This study examines the factors that make standardizing observational health data to the Observational Medical Outcomes Partnership (OMOP) CDM successful.

Materials and methods: Twenty-five data partners (DPs) from 11 countries received funding from the European Health Data Evidence Network (EHDEN) to standardize their data. Three surveys, DataQualityDashboard results, and statistics from the conversion process were analyzed qualitatively and quantitatively. Our measures of success were the total number of days to transform source data into the OMOP CDM and participation in network research.

Results: The health data converted to CDM represented more than 133 million patients. 100%, 88%, and 84% of DPs took Surveys 1, 2, and 3. The median duration of the 6 key extract, transform, and load (ETL) processes ranged from 4 to 115 days. Of the 25 DPs, 21 DPs were considered applicable for analysis of which 52% standardized their data on time, and 48% participated in an international collaborative study.

Discussion: This study shows that the consistent workflow used by EHDEN proves appropriate to support the successful standardization of observational data across Europe. Over the 25 successful transformations, we confirmed that getting the right people for the ETL is critical and vocabulary mapping requires specific expertise and support of tools. Additionally, we learned that teams that proactively prepared for data governance issues were able to avoid considerable delays improving their ability to finish on time.

Conclusion: This study provides guidance for future DPs to standardize to the OMOP CDM and participate in distributed networks. We demonstrate that the Observational Health Data Sciences and Informatics community must continue to evaluate and provide guidance and support for what ultimately develops the backbone of how community members generate evidence.

Keywords: OMOP common data model; data standardization; observational data.

INTRODUCTION

Health data comes in many forms: electronic health records (EHRs) (eg, general practitioner [GP]), clinical registries, longitudinal survey data, insurance claims data, and much more. Health data is often collected and stored in different ways which makes standardized research across data from multiple institutions difficult. Research can be challenging even within a single institution where multiple data sources are used. Conversion of health data to a common data model (CDM) facilitates research by transforming the data into a common format with a standardized vocabulary. This standardization allows for systematic analysis across disparate health data sources.[1] The Observational Health Data Sciences and Informatics (OHDSI) [2] community, a multi-stakeholder, interdisciplinary collaborative that strives to bring value out of health data through large-scale analytics, relies on its CDM model called Observational Medical Outcomes Partnership (OMOP) CDM.[3] This person-centric model is the backbone of how OHDSI improves observational research to produce a comprehensive understanding of health and disease.[4]

The process of standardizing health data to the OMOP CDM is referred to as the extract, transform, and load (ETL) process. The Book of OHDSI [4] suggests 4 main steps for this process: data and CDM experts design the ETL, individuals with medical knowledge map source vocabulary codes to standardized codes, a technical person implements the ETL, and a quality control process is implemented. The OHDSI community created open-source tools for these steps.[5–9] Building and performing an ETL to the OMOP CDM is an investment, but the return on that investment is ready access to sophisticated analytics and ability to participate in network studies.

Even with process recommendations and supporting tools, building an ETL for any CDM can be challenging for some teams. Ong et al. [2] identified 24 technical hurdles that often arise throughout an ETL process; including challenges working with source data, technical difficulties, issues with knowledge management, code management and versioning issues, data quality concerns, and ETL operation challenges. Improving the ETL process starts with understanding these problems. While the OHDSI community has recommendations for developing an ETL, the challenges of developing one have never been formally evaluated.

During the coronavirus disease pandemic (COVID-19) crisis, the European Health Data & Evidence Network (EHDEN), [10] a public-private partnership with a goal to build a large-scale, federated network of health data standardized to the OMOP CDM, held a call for data. The COVID-19 Rapid Collaboration Call, or EHDEN data call, invited institutions to apply for financial and technical support to standardize their data that included COVID-19 patients.[11] The goal was to produce high-quality standardized data to support important characterizations of patients with COVID-19, learn how to best manage their

care, and ensure their treatments are safe and effective.[11] This work is of particular importance in Europe, as converting to the OMOP CDM helps address the most important challenge in international projects, interoperability of different data sources and the difference in terminology, by creating a research environment leveraging federated data sources. Twenty-five data partners (DPs) were awarded the grant. This EHDEN data call provided a unique opportunity to understand key success factors for the development of an ETL, where success can be defined as both a timely development of an ETL and network research involvement.

This work aims to evaluate the critical factors that contribute to the success of standardizing health data to the OMOP CDM. Success was evaluated based on the timeliness of developing the ETL and the ability to participate in network research. To capture these measurements, from each DP, we collected data about their journey to the OMOP CDM as well as surveyed them about any ETL challenges they faced. By understanding what factors lead to a successful transformation of a data source to the OMOP CDM, we will be able to provide further recommendations for a preferred ETL process.

METHODS

COVID-19 Rapid Collaboration Call

Twenty-five DPs representing 11 different countries were awarded financial support to standardize their data to the OMOP CDM under the EHDEN data call.[11] The 11 countries included Belgium, Denmark, Estonia, France, Italy, Portugal, Serbia, Spain, the Netherlands, Turkey, and the United Kingdom. The size of the databases ranged from 400 up to 39 million persons, representing a range of different data types including hospital (inpatient only), EHRs (mix of inpatient and outpatient data), claims, and registry data. Details about each database can be found in Appendix S1 and was sourced from the publicly available EHDEN Catalog (https://portal.ehden.eu/).

Each DP was expected to follow the current OMOP CDM ETL development process suggested by the OHDSI community, as seen in Figure 1A. They started by summarizing their source data using a tool called White Rabbit.[5,12] The output of this step allowed the teams to learn about their data as well as use a tool called Rabbit-in-a-Hat, a graphical user interface facilitating the collaborative design of the ETL. In some cases, the Usagi tool was used to facilitate the mapping of source vocabularies to standard terminologies.[6,13] Once this work was done, the DP was ready to implement their ETL. In this step, DPs chose the tools and methods best suited to their institution to execute the ETL. Once the data was transformed, the resulting database was evaluated using the ACHILLES and DataQualityDashboard (DQD) tools.[7,8] Any issues discovered using these tools were addressed and incorporated back either into the ETL or the vocabulary mapping steps.

The ETL process was re-executed and re-evaluated for quality. The goal of this iterative process was to produce a research-ready CDM database. These tools and processes are described in more detail in the Book of OHDSI.[4]

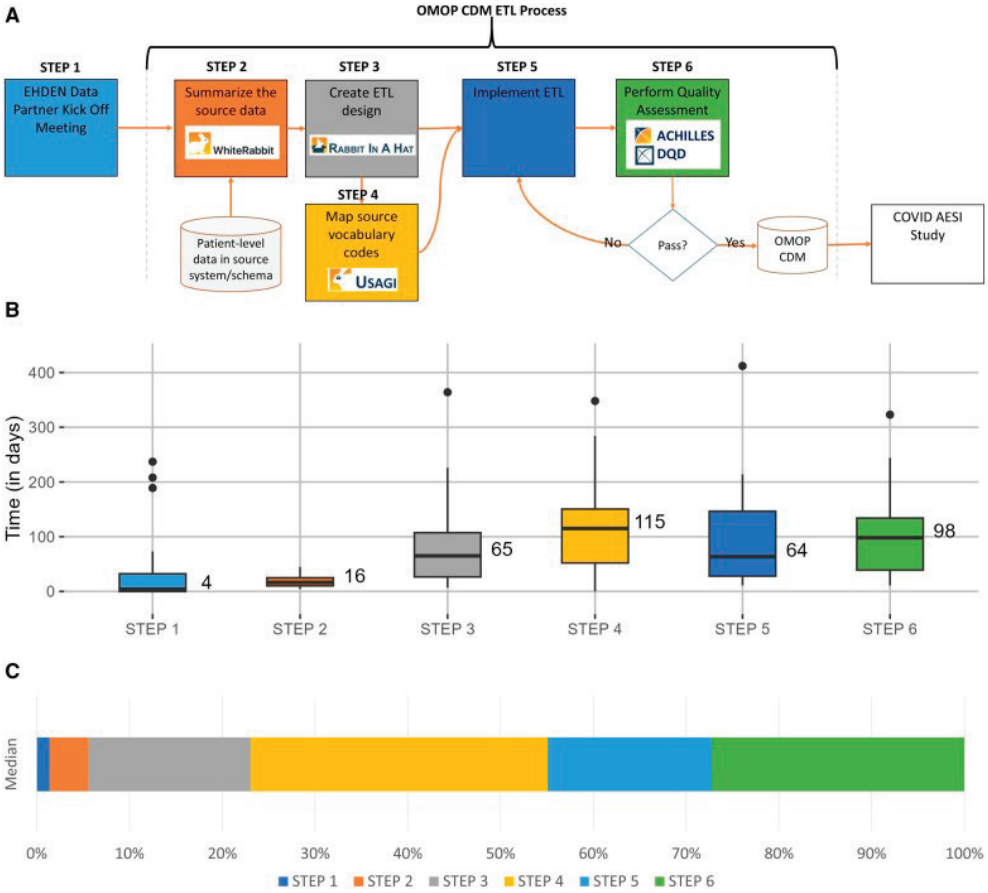


Figure 1: OMOP CDM ETL development process: (A) represents the ETL process map, (B) is a box plot of median length in days for each step across all data partners, and (C) is a stacked bar chart showing the percentage of median time each step took. CDM, common data model; COVID AESI Study, "Adverse Events of Special Interest within COVID-19 Subjects" study; DQD, DataQualityDashboard; EHDEN, European Health Data & Evidence Network; ETL, extract, transform, and load; OMOP, outcomes partnership common data model.

The EHDEN data call diverged from the regular OMOP ETL process in 3 ways. First, each DP, except one, was paired with the EHDEN Taskforce to facilitate the conversion; the EHDEN Taskforce were ETL specialists. The excluded DP knew the OMOP ETL procedure and did not need support. Second, once DPs had completed their ETLs they were offered the opportunity to participate in research initiatives. Specifically, some of the DPs participated in the study "Adverse Events of Special Interest (AESI) within COVID-19 Subjects," or the

COVID-19 AESI study for short, designed to estimate background rates of adverse events after COVID-19 disease.[14–16] This allowed DPs to exercise their data in their new OMOP CDM and obtain network-based research skills. The third and final significant change was that the DPs were requested to complete surveys at various points in the ETL process, allowing us to collect data that might assist in better understanding the critical success elements.

Survey development

Three surveys, found in Appendix S2, were created with a mix of open-ended and multiple-choice questions. The survey questions were developed with the Framework for ETL Challenge Classification outlined by Ong et al.[2] The themes of the framework were divided across the surveys: Survey 1 covered source data; Survey 2 covered technical difficulties and knowledge management; and Survey 3 covered code management and versioning, data quality, and ETL operations. The survey questions were refined based on feedback from 4 individuals who had experience both with the OMOP CDM and EHDEN. Google Forms (https://docs.google.com/forms), a free survey tool, was used to collect responses. Every member associated with the DP team was able to complete the survey, meaning it was possible for multiple replies per survey for each DP.

Data collection

Four types of data were collected. First, the 3 surveys were given as DPs progressed through their ETL. Survey 1 was given after the project kickoff to each DP, Survey 2 was given after the ETL design was completed, and Survey 3 was given at the completion of the OMOP CDM. Survey 1 was taken for the first time on May 13, 2020, and Survey 3 was taken for the last time on May 17, 2022. Since DPs could start and stop their projects at their own rate, progress was fluid during those 2 years.

Data associated with the ETL progress were also collected during regular meetings with DPs. For example, the start and stop of each step were tracked, as illustrated in Figure 1A, notes from meetings, as well as a list of vocabularies found within each data source.

The third type of data collected was the results from running DQD. This was already part of the process each DP needed to go through to complete their work in the EHDEN data call. The number of checks runs in DQD and the outstanding issues were reviewed by the EHDEN Taskforce.

Finally, it was recorded whether the DP participated in the *COVID-19 AESI study*.[16]

Analysis

Our measures of success were the total number of days to transform source data into the OMOP CDM and if a DP participated successfully in network research. DPs were successful if they took less than 365 days to complete their transformation and were part of the *COVID-19 AESI study*.[16] The duration of 365 days was selected because the EHDEN contract stipulated that work should finish within that time. Additionally, in order to understand the association between our key measures of success and the survey responses, we reported the results as bar charts dichotomized by the outcome measures. Depending on the survey question, answers were either summarized as max per DP (max was chosen as a convenient and consistent way to summarize) or summarized across all answers provided by members of the DP. The summary tactics used are discussed in the "Results" section.

RESULTS

In total, 25 DPs participated in this EHDEN data call, representing 11 different countries, collectively covering more than 67 million patient records from claims, EHR, and registries from GP, secondary care, and hospital systems. This represents the largest cross-sectional view of European data to date and is only 13% of the DPs participating in the EHDEN consortium (as of April 2023). Information about the 25 DP's databases can be found in Table 1 with details in Appendix S1.

Table 1: COVID-19 Rapid Collaboration Call Data Partners Information with Participation in EHDEN Taskforce, the Three Surveys, and Network Research Study as of August 2022

Short Name	Full Name	Patient Count	Country	Type	Task-force	Survey			Completed Call for Data	Participated in Study
						1	2	3		
APHM	Health Data Warehouse of Assistance Publique - Hopitaux de Marseille	2.47M	France	Hospital data	Yes	Yes	Yes	Yes	Yes	Yes
APUM *	Azienda Policlinico Universitaria di Modena	400	Italy	EHR	Yes	Yes	Yes	No	No	No
AUMC	Pacmed Data Warehouse at Amsterdam University Medical Center	1.83k	The Netherlands	Hospital data	Yes	Yes	Yes	Yes	Yes	No
BIOCRUCES	Biocruces Bizkaia Health Research Institute	45k	Spain	EHR	Yes	Yes	Yes	Yes	Yes	No
CC_NIS	University Clinical Center Nis	223k	Serbia	EHR	Yes	Yes	No	Yes	Yes	No
CC_SERBIA	University Clinical Center of Serbia	823k	Serbia	EHR	Yes	Yes	Yes	Yes	Yes	Yes
CPRD_AURUM	Clinical Practice Research Datalink - AURUM + Hospital Episode Statistics Admitted Patient Care (HES APC) data	39M	United Kingdom	GP	Yes	Yes	Yes	Yes	Yes	Yes
CSS	Center for Surgical Science	76.8k	Denmark	Registry	Yes	Yes	Yes	Yes	Yes	No
DATALOCH *	DataLoch	414k	United Kingdom	EHR	Yes	Yes	Yes	Yes	Yes	No
FIIBAP *	Fundación para la Investigación e Innovación Biosanitaria en Atención Primaria	292k	Spain	EHR	Yes	Yes	Yes	Yes	Yes	Yes
FINCB *	Fondazione IRCCS Istituto Neurologico Carlo Besta	766	Italy	Hospital data	Yes	Yes	Yes	Yes	Yes	No
FPIO *	Fondazione Poliambulanza Istituto Ospedaliero	23.1k	Italy	EHR	Yes	Yes	Yes	Yes	Yes	No
HDH *	Health Data Hub	91.9k	France	EHR	Yes	Yes	Yes	No	No	No
HIC	Health Informatics Centre	1.3M	United Kingdom	EHR	Yes	Yes	Yes	Yes	Yes	Yes
IDIVAL	Servicio Cántabro de Salud and IDIVAL	580k	Spain	EHR	Yes	Yes	Yes	Yes	Yes	No
IMASIS	Parc de Salut Mar Barcelona Information System (IMASIS)	976k	Spain	EHR	Yes	Yes	Yes	Yes	Yes	Yes
IRCCSE *	Azienda Unità Sanitaria Locale - IRCCS in Reggio Emilia	1.8k	Italy	EHR	Yes	Yes	Yes	Yes	Yes	No
IU	Istanbul Faculty of Medicine, Istanbul University	899k	Turkey	EHR	Yes	Yes	No	Yes	Yes	Yes
LYNXCARE	LynxCare Clinical Informatics	4 dbs (range 2.5k-20k)	Belgium	EHR	Yes	Yes	Yes	Yes	Yes	No
MEDAMAN	Medaman Hospital Data	117k	Belgium	EHR	Yes	Yes	Yes	Yes	Yes	Yes
RCGP	Royal College of General Practitioners Research and Surveillance Centre	11M	United Kingdom	EHR	Yes	Yes	Yes	No	No	No
SIDIAP	The Information System for Research in Primary Care	8M	Spain	GP	Yes	Yes	Yes	Yes	Yes	Yes
U_OF_TARTU	University of Tartu	386k	Estonia	Claims	No	Yes	No	No	Yes	Yes
UK_BIOBANK	UK Biobank	502k	United Kingdom	Registry + EHR	Yes	Yes	Yes	Yes	Yes	Yes
ULSM *	Unidade Local de Saúde de Matosinhos	9.75k	Portugal	EHR	Yes	Yes	Yes	Yes	Yes	No

* = COVID-19 Only Data sets; dbs = databases; EHR = electronic health record; GP = general practitioner

Survey results

As of August 2022, all 25 DPs completed Survey 1, while 88% completed Survey 2, and 84% Survey 3. Survey 1 had an average of 4.6 individuals from each DP respond (1 minimum—11 maximum). Both Survey 2 and Survey 3 had an average of 1.7 individuals from each DP respond (1 minimum—4 maximum). Twenty-two DPs completed all requirements of the call. Twenty-one DPs received support from the EHDEN Taskforce. Ultimately 21 DPs both completed the EHDEN data call and worked with an EHDEN Taskforce. Most of the results are reported for the 21 DPs and all results are specified if they are for the 21 DPs or for all 25 DPs. Ten (48%) of the 21 DPs participated in the COVID-19 AESI Study.

Table 2 summarizes key questions from Survey 1 for 21 DPs. The first question summarized was "What will be your primary role in this project?" For example, 20 out of 21 DPs had at least one person whose primary role in the project was "informatician." Another question was "How would you classify your expertise with the data source?" Eleven DPs had at least one person who considered themselves an expert. Finally, for the question "Realistically, how many hours a week can you dedicate to this project?," 6 DPs had at least one person planning on spending 9-16 h/week on the EHDEN data call.

Table 3 summarizes key questions from Survey 2 for the 21 of which 90% responded. When the DPs were asked "How many tables from your source data are in your ETL?" (max number by DP), 29% reported having less than 10 tables, 33% reported having 10 or more tables, and 38% did not reply. When asked "Thinking of your source data's tables being converted to the OMOP CDM, prior to starting this Rapid Collaboration Call how much experience did you or your team have with these tables needed in the ETL?" (max choice of most familiar selected by DP), 52% reported being familiar with most or all of the tables, 38% reported being familiar with none, few, or some of the tables, and 10% did not reply. When asked "To use your source data in this Rapid Collaboration Call was there any effort necessary to prepare the data?" (max choice of most effort selected by DP), all DPs that replied to the survey said there was some effort to prepare the data (38% said there was much effort and 52% said there was some effort needed). When asked "Prior to the Rapid Collaboration call were you familiar with the OMOP Common Data Model?" (max choice of most familiar selected by DP), 57% said they were familiar. Finally, when asked, "Have you learned anything about your source data by going through this process?" (max choice of someone learning was selected by DP), 62% responded "Yes," 24% responded "No," 5% responded "Based on my involvement I cannot comment.," and 10% did not answer.

Table 2: Summary Of Key Questions In Survey 1 For 21 Data Partners (that were both Completed with the COVID-19 Rapid Collaboration Call and Worked with the EHDEN Taskforce)

Survey Question	Options	# of Data Partners
What will be your primary role in this project? *(summarized as often as the role showed up, it is possible for multiple roles to be present on one team, however the role was only counted once per team)*	Informatician	20
	Computer scientist	17
	Project manager	16
	Data manager	8
	Clinical scientist	7
	Person in Medicine	4
	Health Policy Individual	3
	Epidemiologist	3
	Statistician	2
	Something Other Than the Above	5
	Preferred Not to Say	0
How would you classify your expertise with the data source? *(summarized max per data partner, every data partner counted once)*	Novice (minimal knowledge of the data source)	0
	Beginner (working knowledge of the data source)	2
	Competent (good working knowledge of the data source)	5
	Proficient (in depth understanding of the data source)	3
	Expert (authoritative knowledge of data source)	11
Realistically, how many hours a week can you dedicate to this project? *(summarized max per data partner, every data partner counted once)*	0-4 hours / week	2
	5-8 hours / week	5
	9-16 hours / week	6
	17-24 hours / week	3
	25-32 hours / week	2
	33-40 hours / week	3

Table 3: Summary of Key Questions in Survey 2 for 21 Data Partners (that were both Completed with the COVID-19 Rapid Collaboration Call and Worked with the EHDEN Taskforce)

Survey Question	Options	# of Data Partners
How many tables from your source data are in your ETL? *(selected the max number reported in survey, these numbers were then categorized into <10 tables and >=10 tables)*	Less than 10 tables	6
	Greater than or equal to 10 tables	7
	No response reported	8
Thinking of your source data's tables being converted to the OMOP CDM, prior to starting this Rapid Collaboration Call how much experience did you or your team have with these tables needed in the ETL? *(selected the max choice from data partner, the choices were dichotomized into two options)*	Familiar with none / few / some of the tables	8
	Familiar with most / all of the tables	11
	No response reported	2

Table 3: Summary of Key Questions in Survey 2 for 21 Data Partners (that were both Completed with the COVID-19 Rapid Collaboration Call and Worked with the EHDEN Taskforce) (*continued*)

Survey Question	Options	# of Data Partners
To use your source data in this Rapid Collaboration Call was there any effort necessary to prepare the data? For example, did a data extract from source systems need to be put in place in order for data to be available for the ETL process? *(selected the max choice per data partner)*	Yes, much effort was needed to prepare the raw data for ETL.	8
	Yes, some effort was needed to prepare the raw data for ETL.	11
	No, the data was in a format prior to applying to the Rapid Collaboration Call that was suitable for ETL.	0
	No response reported	2
Prior to the Rapid Collaboration call were you familiar with the OMOP Common Data Model? *(selected the max choice per data partner)*	Yes	12
	No	7
	No response reported	2
Have you learned anything about your source data by going through this process?	Yes	13
	No	5
	Based on my involvement I cannot comment.	1
	No response reported	2

Table 4 summarizes key questions from Survey 3 for 21 DPs of which all had at least one response. When asked about the complexity of the ETL process (max level of difficulty reported by DP), 67% of DPs found the process easy or neutral in complexity. When asked about the effectiveness of the tools used to assess data quality (eg, ACHILLES, DQD) (selecting the least helpful choice by DP), 71% of DPs found the tools supportive but needed support from the EHDEN Taskforce to use them appropriately. Additionally, only 19% of DPs said that their organization had formal plans to improve their CDM and use it for research in the future, 57% said the organization still needs to better understand the value to continue moving forward, and 24% were not sure of the plans for the CDM moving forward (max reply of how confident the DP was that the CDM would be used was selected). Finally, the ETL step that was most challenging for DPs was mapping source vocabulary codes to standardized concepts (all unique answers were summarized).

ETL process measures

Of the 21 DPs, the median time it took to complete the ETL process was 358 days, with the shortest time being 172 days and the longest being 622 days and an interquartile range of 276-481. Figure 1B illustrates the duration of each step in the ETL process. Figure 1C is a stacked bar chart illustrating which process steps consumed the most time (based on the median days). Step 4 (mapping vocabulary codes) and Step 6 (performing quality assessment) required the greatest time. To understand the reasons why individual steps took the time they did, the regular meeting notes were used to understand the details.

In Step 1, 1 DP required 237 days to begin working with their data, 3 DPs required more than 100 days, and 9 DPs experienced some delay. The cause for the delay was almost invariably a lack of appropriate staff or data access issues. Step 2 was the least complicated of the steps; it consisted of a WhiteRabbit database scan; however, 1 DP required 45 days to complete the assignment. This was because this DP's data was dispersed over multiple systems, and it required some time to find the appropriate personnel to assist with the scan. Step 3, creating the ETL, took 1 DP 364 days to complete. The same DP also took the longest to complete Step 5, completing data quality evaluation, with 412 days. These processes took a long time since the project's primary developer did not always have the required access or rights to complete the task, resulting in several technical obstacles over the duration of the project. In addition, the person most knowledgeable about the data was not always accessible to the lead developer; therefore, we frequently had to wait for the two to communicate in order to resolve issues. Step 4, vocabulary mapping, took 1 DP 348 days to complete due to the rearrangement of the team working on the process, which resulted in little to no work being performed for the majority of the time. Once the new team was assigned, work went rapidly. Many of the DPs that took the longest at each phase were affected by the availability of the appropriate personnel and access or readiness of data.

Table 4: Summary of Key Questions in Survey 3 for 21 Data Partners (that were both Completed with the COVID-19 Rapid Collaboration Call and Worked with the EHDEN Taskforce)

Survey Question	Options	# of Data Partners
Given now that you are at the end or near the end of this process, how complex did you find the ETL process was for your data source to be converted to the OMOP Common Data Model? *(selected the max choice from data partner, the choices were dichotomized into two options)*	01) Difficult	7
	02) Easy & Neutral	14
There are several tools we used to assess data quality (i.e. ACHILLES, Data Quality Dashboard, CDM Inspection Report, and Catalog Export). Please select the answer that best fits your experience: *(selected the max choice per data partner)*	Even with the tools, I did not feel like there were detailed specifications for what to assess. Even with the help from the EHDEN Task Force I did not find the tools helpful or informative.	0
	The tools were supportive, however without the EHDEN Task Force I am not sure if I would have been able to make sense of what to do with the information.	6
	The tools provided insight, and I would have been able to make improvements on my own (without the EHDEN Task Force) but it would have been difficult or slow.	9
	The tools were useful, with them alone I could have made most of the necessary improvements needed for my CDM.	3

Table 4: Summary of Key Questions in Survey 3 for 21 Data Partners (that were both Completed with the COVID-19 Rapid Collaboration Call and Worked with the EHDEN Taskforce) (*continued*)

Survey Question	Options	# of Data Partners
There are several tools we used to assess data quality (i.e. ACHILLES, Data Quality Dashboard, CDM Inspection Report, and Catalog Export). Please select the answer that best fits your experience: *(selected the max choice per data partner)*	The tools provided an obvious way to assess data quality and I was clear how to interpret the results.	2
	Based on my involvement I cannot comment.	1
Please choose the statement that most accurately describes your organization's view on maintaining the CDM moving forward: *(selected the max choice from data partner, the choices were simplified to 3 options)*	01) I am not aware of my organization's thoughts on our CDM's use moving forward.	5
	02) We have made our CDM, have some plans to update, and may still need to see the value.	12
	03) We have made our CDM and plan to improve upon it moving forward to facilitate our organization's research on our medical data.	4
Please select which step in the ETL process you found the most challenging: *(all unique choices per data partner were selected, each response can be up to 21)*	Summarizing the Source Data with White Rabbit	2
	Create ETL Design	2
	Mapping Source Vocabulary Codes to Standardized Concepts	11
	Setting up an environment for the processing of the ETL	2
	Implement ETL	6
	Perform Data Quality Assessment	5
	Other	3

Of the 21 DPs, 20 had at least one source vocabulary that needed to be mapped to standardized terminologies using the OMOP Vocabulary or tools like Usagi. The remaining DP did not spend time mapping vocabularies as they were planning to adopt the same EHR platform as another DP and thus could leverage their mapping work. Among the 20 DPs, the median number of source vocabularies was 7, ranging from 1 to 28. The median number of these vocabularies that were not in the OMOP Vocabulary and required the DP to spend time linking source vocabularies to standard terminology was 3.5, ranging from 1 to 21. This means that DPs needed to map a median of 59% of their source vocabulary concepts, ranging from 13% to 100%.

Results from DQD

DQD results were shared from 20 of the DPs. The failure to collect DQD files from a single DP was an oversight. The median number of times a DP ran the DQD was 3 with an inter-

quartile range of 2-7. Step 6, which coupled DQD and ACHILLES review, took a median of 98 days that ranged between 11 and 323 days. The most common issues identified in the first run were related to how well the database conformed to the technical specifications of the OMOP CDM. As these were addressed, subsequent runs of the DQD revealed more complicated issues, typically related to the mapping of site-specific codes to the standardized vocabulary and vocabulary domain harmonization.[17]

Success measures

Of the 21 DPs, 52% had built their CDM in under 365 days, 48% participated in the COVID-19 AESI study,16 and 38% did both. A few correlations can be seen between the results and success markers. For instance, when evaluating how long it took a DP to start working on the project, Figure 2 demonstrates that 63% of those DPs who started right away (ie, the same day) met both success criteria as opposed to 31% of those who did not. Furthermore, DPs who thought the ETL process would be difficult before beginning the ETL were less likely to achieve the success indicators. DPs that did not achieve success 79% thought the process would be difficult versus 57% of DPs who did achieve success and thought the process would be difficult (Figure 3A). This trend persisted even after the process was completed (46% of the DPs who failed to meet the success indicators found the ETL process difficult, but only 13% of the DPs who met the success criteria did) (Figure 3B).

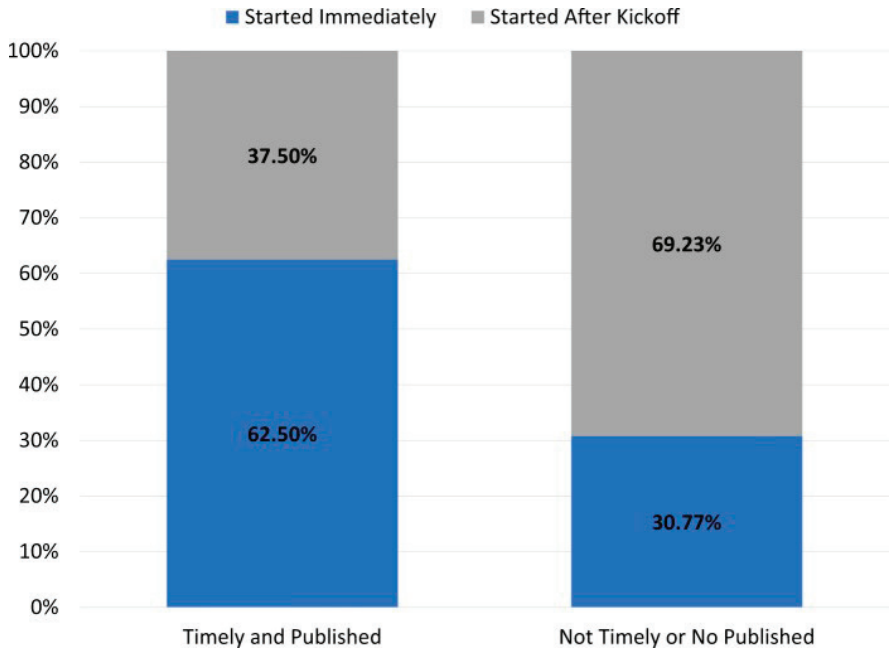


Figure 2: For data partners that met success criteria (timely and published, n = 8) versus those who did not meet the success criteria (not timely or not published, n = 13), how often did that data partner start immediately or after kick off.

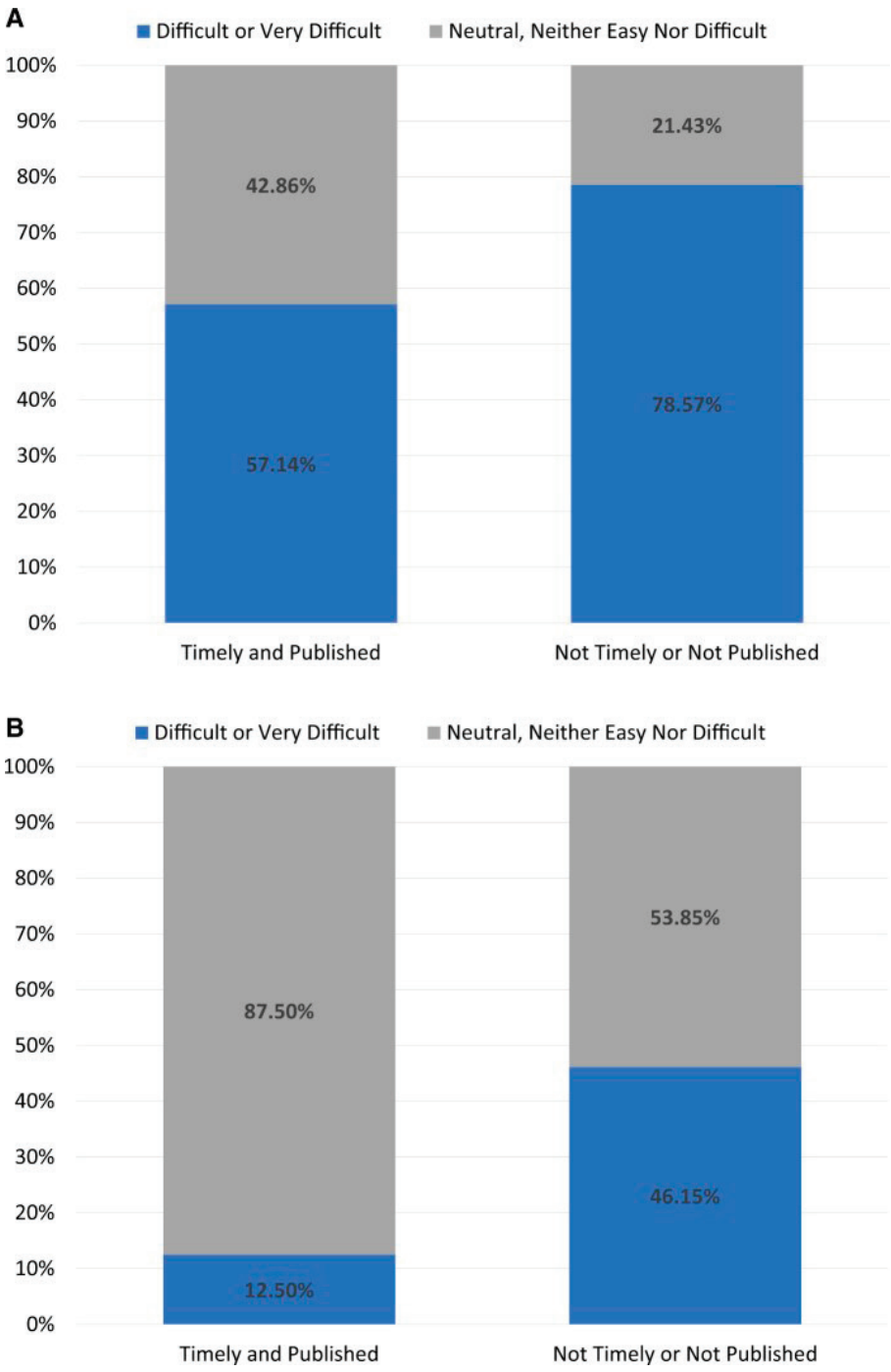


Figure 3: For data partners that met success criteria (timely and published, n = 8) versus those who did not meet the success criteria (not timely or not published, n = 13): (A) represents at the beginning of the process how difficult did they think the process would be and (B) represents at the end of the process how often did that data partner find the process difficult.

DISCUSSION

To the best of our knowledge, this is the most comprehensive assessment of the critical factors that contribute to the successful conversion of complex health-care data to the OMOP CDM. It was performed during the COVID-19 pandemic, a period when access to data was limited and team members were not continuously available. Nevertheless, most ETLs where realized within the agreed timelines. By doing this, all DPs were able to successfully improve the interoperability of their data which allowed them to participate in research studies using standardized analyses.

Our study identified multiple factors that had a major impact on timelines which were uncovered using the ETL measures captured. For example, studying the duration of tasks and reviewing themes among DPs that took a long time on certain steps allowed us to uncover the importance of establishing governance rules prior to the initiation of the work. Also, through both the surveys and ETL measures captured, we learned the right composition of the team proved to be very important. The team needs to contain members with deep knowledge on the data source, members that have a good understanding of the OMOP CDM and the vocabularies, and experts to implement the ETL.

Improving the interoperability of health data requires standardizing both the structure (syntactic interoperability) and the terminologies (semantic interoperability). The surveys revealed that mapping from the source structure to the standardized clinical tables of the OMOP CDM was not seen as difficult by the teams but mapping source codes to the OHDSI Standardized Vocabularies was frequently mentioned as a challenging aspect of the ETL process. This is aligned with previous literature on ETL development.[18–32] For example, a recent publication from Oja et al.[18] highlighted the difficulty in selecting the correct concept when there are multiple options, and issues around different levels of granularity when mapping source codes to target codes. Vocabulary mappings require medical expertise and in-depth knowledge about both the source vocabulary and the target vocabulary. Also, knowing when to stop mapping terminology is key to not wasting time, often a select subset of terms will make up the majority of the database record thus all terms do not need to be mapped.[29] There is clear value in additional training on the mapping strategy and compliance with the OMOP conventions.

Furthermore, the vocabulary mapping process needs to be supported by tools and quality control steps. The Usagi mapping tool developed by the OHDSI community was much appreciated by the teams but there is room for further improvement, eg, to add support for collaborative mapping in which multiple members can review and approve results. As a response to the lessons learned from this study, the EHDEN consortium and OHDSI community have recently started work to improve vocabulary management, quality control, and mapping. EHDEN, for example, has implemented collaboration functionality in

Usagi and OHDSI collaborators are developing new tools to support vocabulary mapping (eg, Perseus,[33] Susana[34]).

Due to the pandemic, the ETL was supported by the EHDEN Taskforce through many online meetings. These meetings were where most of the ETL process measures were captured and were important to provide guidance on ETL design and source vocabulary mappings. Clearly, a more effective approach is to hold a multi-day face-to-face meeting with all the team members. Based on past experience this would have allowed us to design the ETLs in a much shorter time (eg, the Integrated Primary Care Information database ETL was designed in 1 month[35]).

Implementing an ETL also requires preparation in terms of data access and personnel resources. Data access delays were very common and captured through the ETL process measures. For 1 DP, it took over 6 months to get institutional review board approval to standardize their data. No progress could be made before data access was allowed. Furthermore, some DPs required a considerable amount of time to get started with the ETL design and implementation due to challenges in forming an appropriate team. For example, one team was made up entirely of physicians with no technical background, making it extremely challenging to initiate a technical task such as developing an ETL. Once a technical expert was added to the team considerable progress was made. Availability of team members was limited especially for those institutions that were severely impacted by COVID-19. The support network of certified Small to Medium-sized Enterprise (SMEs) created by the EHDEN project could be a good alternative for organizations that do not have all competences in house.

Previous papers discussing ETL transformation[1,18–20,23,28,29,32] have seen the need for appropriate resources and appropriate access to data. For example, Overhage et al.[32] saw the need for having the appropriate people and access to data and stated, "Each partner utilized a number of people with a wide range of expertise and skills to complete the project, including project managers, medical informaticists, epidemiologists, database administrators, database developers, system analysts/programmers, research assistants, statisticians, and hardware technicians." Similar recommendations were made by other studies,[1,18,20,23,28,29] and a similar recommendation can even be found in the Book of OHDSI.[4] Having sufficient access to data before beginning ETL construction, however, has received less attention in prior research. Only 2 recent publications, Oja et al.[18] and Yu et al.,[19] discuss obtaining permissions to utilize the data. As we have shown, obtaining correct approval can affect the duration of the ETL development and should be considered early on.

In this work, we successfully transformed 25 different databases into the OMOP CDM using a standard process and toolkit. Despite different geographies, data types, source vocabularies, and population sizes, as well as different team compositions with differing expertise, the consistent workflow used by EHDEN proves appropriate to support the successful standardization of observational data across Europe. Across the 25 successful transformations, we continue to solidify the notions that having the appropriate persons present for the ETL and vocabulary mapping can be a challenging aspect, and in addition, we learned that groups should proactively prepare for data governance issues. This effort should provide guidance for future DPs to standardize to the OMOP CDM and participate in distributed networks. The OHDSI community must continue to evaluate and provide guidance and support for what ultimately develops the backbone of how community members generate evidence.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Journal of the American Medical Informatics Association online.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10746315/#sup1

ACKNOWLEDGEMENTS

We would like to acknowledge the COVID-19 Rapid Collaboration Call Data Partners that supported this work. We would like to specifically thank our DPs at HDH (Lorien Benda, Cécile Charles, Gaëlle Rimaud, Axelle Menu-Branthomme, Emmanuel Stranadica, Anne Cuerq, and Tim Vlaar), at Biocruces (Eunate Arana-Arri and Iñaki Gutiérrez-Ibarluzea), at IRCCS (Romina Corsini), at IU (Selçuk Şen and Ali Yağız Üresin), at UCC_NIS (Dimitrije Spasić, Jelena Miladinović, and Milan Lazarević), at ULSM (Rosa Princípe, José Castanheira, Tiago Morais, Daniel Santos, and Tiago Taveira-Gomes), at Hospital del Mar with the IMASIS database (Juan Manuel Ramírez-Anguita and Miguel A. Mayer), at IDIAPJGol with the SIDIAP database (Sergio Fernández-Bertolín, Maria Aragón, and Talita Duarte-Salles), at MDM (Luc Belmans, Geert Byttebier, and Manu L.N.G. Malbrain), and at CSS (Andi Tsouchnika and Andreas Weinberger Rosen). We would also like to acknowledge the patients and healthcare professionals who suffered the consequences of the COVID-19 pandemic, and whose data was fundamental to improve our understanding of the determinants and outcomes of this novel disease.

CONTRIBUTOR INFORMATION

Erica A Voss, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands. Janssen Pharmaceutical Research and Development LLC, Raritan, NJ 08869, United States.

Clair Blacketer, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands. Janssen Pharmaceutical Research and Development LLC, Raritan, NJ 08869, United States.

Sebastiaan van Sandijk, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Odysseus Data Services, Prague, Czech Republic.

Maxim Moinat, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands.

Michael Kallfelz, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Odysseus Data Services, Prague, Czech Republic.

Michel van Speybroeck, Janssen Pharmaceutical Research and Development LLC, Raritan, NJ 08869, United States.

Daniel Prieto-Alhambra, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands. Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, United Kingdom.

Martijn J Schuemie, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Janssen Pharmaceutical Research and Development LLC, Raritan, NJ 08869, United States. Department of Biostatistics, University of California, Los Angeles, CA 90095, United States.

AUTHORS CONTRIBUTIONS

Contributions of co-authors are described following the CRediT, Contributor Roles Taxonomy (https://credit.niso.org/). For conceptualization, or formulation of the research goals, the following authors participated: E.A.V., C.B., M.J.S., and P.R.R. For data curation, or those who worked with research data, the following authors participated: E.A.V., C.B., M.M., M.K., and M.v.S. For formal analysis, or those who analyzed or synthesized study data, the following authors participated: E.A.V. and C.B. For investigation or conducting of the research or data/evidence collection, the following authors participated: E.A.V. Design of the methodology was performed by the following individuals: E.A.V., C.B., M.J.S., and P.R.R. The following individuals performed project administration: E.A.V. Resources, or provision of computing resources or analysis tools, was performed by the following: E.A.V. and P.R.R. Software, or program development, was performed by: E.A.V., C.B., S.v.S., M.M., MC, and P.R.R. Supervision, or oversight, was performed by: D.P.-A., M.J.S., and P.R.R. Validation, or verification of results, was performed by: C.B., S.v.S., M.M., and M.v.S. Visualization, or creation of visualizations or data presentation, was performed by: E.A.V. and C.B. Writing of the original draft was performed by: E.A.V. and C.B. Editing and reviewing the original writing was performed by: E.A.V., C.B., S.v.S., M.M., M.K., M.v.S., D.P.-A., M.J.S., and P.R.R. All authors approved the final version, had access to the data, and accept responsibility to submit the final version for publication. No competing interests.

Peter R Rijnbeek, OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICTS OF INTEREST

E.A.V., C.B., M.v.S., and M.J.S. are employees of Janssen Research and Development LLC and shareholders of Johnson & Johnson (J&J) stock. P.R.R. on this project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking (JU) under grant agreement number 806968. The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. The funding source had no role in writing the manuscript or the decision to submit it for publication. P.R.R. works for a department that receives/received unconditional research grants from UCB, Chiesi, Amgen, Johnson and Johnson, European Medicines Agency. Odysseus Data Services, where M.K. and S.v.S. are employed, is a beneficiary of the mentioned IMI 2 JU project. D.P.-A.'s research group has received funding from the European Medicines Agency and Innovative Medicines Initiative. D.P.-A.'s department has received grant(s) from Amgen, Chiesi-Taylor, Lilly, Janssen, Novartis, and UCB Biopharma. His research group has received consultancy fees from AstraZeneca and UCB Biopharma. Amgen, Astellas, Janssen, Synapse Management Partners, and UCB Biopharma have funded or supported training programs organized by D.P.-A.'s department.

DATA AVAILABILITY

The data described above will be shared on reasonable request to the corresponding author.

REFERENCES

1. Reisinger SJ, Ryan PB, O'Hara DJ, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010;17(6):652-662. 10.1136/jamia.2009.002477
2. Ong T, Pradhananga R, Holve E, Kahn MG.. A framework for classification of electronic health data extraction-transformation-loading challenges in data network participation. EGEMS (Washington, DC). 2017;5(1):10. 10.5334/egems.222
3. Common Data Model (CDM). OMOP Common Data Model (CDM). Accessed May 14, 2023. https://github.com/OHDSI/CommonDataModel
4. Observational Health Data Sciences and Informatics. The Book of OHDSI. The Book of OHDSI. Accessed August 29, 2019. https://ohdsi.github.io/TheBookOfOhdsi/
5. OHDSI/WhiteRabbit [program]. Version V0.10.1. GitHub. Accessed May 14, 2023. https://github.com/OHDSI/WhiteRabbit
6. OHDSI/Usagi [program]. Version V1.3.0. GitHub. Accessed May 14, 2023. https://github.com/OHDSI/Usagi
7. OHDSI/Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) [program]. Version V1.6.3. GitHub. Accessed May 14, 2023. https://github.com/OHDSI/Achilles
8. OHDSI/DataQualityDashboard (DQD) [program]. GitHub. Accessed May 14, 2023. https://github.com/OHDSI/DataQualityDashboard
9. OHDSI Forums. OHDSI Forums [Web Forum]. Accessed May 14, 2023. https://forums.ohdsi.org/
10. European Health Data & Evidence Network (EHDEN). European Health Data & Evidence Network (EHDEN) [Web Page]. Accessed May 14, 2023. https://www.ehden.eu/
11. European Health Data & Evidence Network (EHDEN). 04/2020—COVID19 Rapid Collaboration Call. 04/2020—COVID19 Rapid Collaboration Call. Accessed November 6, 2022. https://www.ehden.eu/open-calls/04-2020-covid19-data-partner-call/
12. OHDSI White Rabbit. OHDSI White Rabbit. Accessed November 6, 2022. http://ohdsi.github.io/WhiteRabbit/index.html
13. OHDSI Usagi. OHDSI Usagi. Accessed August 27, 2023. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi
14. Adverse Events of Special Interest within COVID-19 Subjects Code Repository. Adverse Events of Special Interest within COVID-19 Subjects Code Repository [GitHub Repository] 2021. Accessed November 6, 2022. https://github.com/ohdsi-studies/Covid19SubjectsAesiIncidenceRate
15. Voss EA, Sa Ostropolets A, Nyberg F, et al. Adverse Events of Special Interest within COVID-19 Subjects—Research Protocol. Adverse Events of Special Interest within COVID-19 Subjects—Research Protocol. Accessed November 6, 2022. https://ohdsi-studies.github.io/Covid19SubjectsAesiIncidenceRate/Protocol.html
16. Voss EA, Shoaibi A, Yin Hui Lai L,. et al. Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 databases: a multinational retrospective cohort study. EClinicalMedicine. 2023;58:101932. 10.1016/j.eclinm.2023.101932
17. Blacketer C, Voss EA, DeFalco F, et al. Using the data quality dashboard to improve the EHDEN network. Appl Sci. 2021;11(24):11920. [Google Scholar]
18. Oja M, Tamm S, Mooses K, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned, medRxiv 2023:2023.02.16.23285697. 10.1101/2023.02.16.23285697, preprint: not peer reviewed.

19. Yu Y, Zong N, Wen A, et al. Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration. J Biomed Inform. 2022;127:104002. 10.1016/j.jbi.2022.104002
20. Puttmann D, De Keizer N, Cornet R, Van Der Zwan E, Bakhshi-Raiez F.. FAIRifying a quality registry using OMOP CDM: challenges and solutions. Stud Health Technol Inform. 2022;294:367-371. 10.3233/shti220476
21. Carus J, Nürnberg S, Ückert F, Schlüter C, Bartels S.. Mapping cancer registry data to the episode domain of the Observational Medical Outcomes Partnership Model (OMOP). Appl Sci. 2022;12(8):4010. [Google Scholar]
22. Papez V, Moinat M, Voss EA, et al. Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. J Am Med Inform Assoc. 2022;30(1):103-111. 10.1093/jamia/ocac203
23. Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol. 2021;21(1):238. 10.1186/s12874-021-01434-3
24. Lamer A, Depas N, Doutreligne M, et al. Transforming French electronic health records into the observational medical outcome partnership's common data model: a feasibility study. Appl Clin Inform. 2020;11(1):13-22. 10.1055/s-0039-3402754
25. Klann JG, Joss MAH, Embree K, Murphy SN.. Data model harmonization for the All Of Us Research Program: transforming i2b2 data into the OMOP Common Data Model. PLoS One. 2019;14(2):e0212463. 10.1371/journal.pone.0212463
26. You SC, Lee S, Cho SY, et al. Conversion of National Health Insurance Service-National Sample Cohort (NHIS-NSC) Database into Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM). Stud Health Technol Inform. 2017;245:467-470.
27. Yoon D, Ahn EK, Park MY, et al. Conversion and data quality assessment of electronic health record data at a Korean Tertiary Teaching Hospital to a common data model for distributed network research. Healthc Inform Res. 2016;22(1):54-58. 10.4258/hir.2016.22.1.54
28. FitzHenry F, Resnic FS, Robbins SL, et al. Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. Appl Clin Inform. 2015;6(3):536-547. 10.4338/aci-2014-12-cr-0121
29. Matcho A, Ryan P, Fife D, Reich C.. Fidelity assessment of a clinical practice research datalink conversion to the OMOP Common Data Model. Drug Saf. 2014;37(11):945-959. 10.1007/s40264-014-0214-3
30. Makadia R, Ryan PB.. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) common data model. EGEMS (Washington, DC). 2014;2(1):1110. 10.13063/2327-9214.1110
31. Zhou X, Murugesan S, Bhullar H, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf. 2013;36(2):119-134. 10.1007/s40264-012-0009-3
32. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE.. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60. 10.1136/amiajnl-2011-000376
33. Perseus. Perseus [Git Repository]. Accessed January 28, 2023. https://github.com/OHDSI/Perseus
34. Susana [Web Page]. Accessed May 14, 2022. https://interhop.frama.io/omop/susana/

de Ridder MAJ, de Wilde M, de Ben C, et al. Data resource profile: the Integrated Primary Care Information (IPCI) database, The Netherlands. Int J Epidemiol. 2022;51(6):e314-e323. 10.1093/ije/dyac026

2

Evaluation of the Impact of Defining Observable Time in Real-World Data on Outcome Incidence

Clair Blacketer, MPH[1,2,4]
Frank J. DeFalco, BA[1,4]
Mitchell M. Conover, PhD[1,4]
Patrick B. Ryan, PhD[1,3,4]
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

Blacketer C, DeFalco FJ, Conover MM, Ryan PB, Schuemie MJ, Rijnbeek PR. Evaluation of the impact of defining observable time in real-world data on outcome incidence. J Am Med Inform Assoc. 2025 Sep 1;32(9):1434-1444. doi: 10.1093/jamia/ocaf119.

ABSTRACT

Objective: In real-world data (RWD), defining the observation period—the time during which a patient is considered observable—is critical for estimating incidence rates and other outcomes. Yet, in the absence of explicit enrollment information, this period must often be inferred, introducing potential bias.

Materials and Methods: This study evaluates methods for defining observation periods and their impact on incidence rate estimates across multiple database types. We applied three methods for defining observation periods: (1) a persistence + surveillance window approach, (2) an age- and gender-adjusted method based on time between healthcare events, and (3) the min/max method. These were tested across 11 RWD databases, including both enrollment-based and encounter-based sources. Enrollment time was used as the reference standard in eligible databases. To assess the impact on epidemiologic results, we replicated a prior study of adverse event incidence, comparing incidence rates and calculating mean squared error (MSE) between methods.

Results: Incidence rates decreased as observation periods lengthened, driven by increases in the person-time denominator. The persistence + surveillance method produced estimates closest to enrollment-based rates when appropriately balanced. The min/max approach yielded inconsistent results, particularly in encounter-based databases, with greater error observed in databases with longer time spans.

Discussion: These findings suggest that assumptions about data completeness and population observability significantly affect incidence estimates. Observation period definitions substantially influence outcome measurement in RWD studies.

Conclusion: Standardized, transparent approaches are necessary to ensure valid, reproducible results—especially in databases lacking defined enrollment.

Keywords: Observation period, real-world data, incidence rates, data standardization, data quality

1 INTRODUCTION

Unlike randomized controlled trials (RCTs) and other prospective studies, which involve direct data collection on patients from registration through follow-up, re-using existing healthcare data presents a unique challenge: it is often unclear when a patient is actually "observed."[1] In real-world data (RWD), an observation period reflects the time during which a patient is expected to have healthcare events recorded. A lack of data during this period may imply either that no clinical events occurred or that there was simply a failure to capture these events. Defining observable time—when we can confidently interpret the absence of data as the absence of clinical events, rather than as missing data—is therefore essential to avoid bias in observational studies[2]. Observable time definitions underpin key outcomes, including incidence rates, time-to-event analyses, and other measures foundational to reliable evidence generation.

In certain data types, such as administrative claims and national insurance-based systems, we have robust information on observable time[3–7]. For example, insurance enrollment records can indicate the periods during which patient interactions with the healthcare system are expected to be captured. However, for other types of data, particularly electronic health records (EHRs), where only encounters with individual sites are recorded, determining observable time is more complex and often requires making assumptions that are difficult to validate[8–12]. A commonly used approach in such cases is the min/max method, which assumes continuous observation between a patient's first and last recorded healthcare event[13,14]. However, this assumption is not foolproof and may overlook gaps in observation, where healthcare events may have occurred in a different office or facility, as well as miss time when a person is observed but not experiencing healthcare events. These assumptions, although critical, are typically made on an ad hoc basis by individual researchers conducting studies and are rarely documented[15–18].

With the growing use of Common Data Models (CDMs), the definition of observable time is becoming increasingly standardized. CDMs enable the sharing of analytical tools and even entire study implementations across different data sources, thereby enhancing transparency and reproducibility[19]. Converting data to a CDM often requires restructuring the data to meet minimum research-readiness standards. For instance, the OMOP CDM necessitates an explicit definition of the observation period, which dictates the span of time in which healthcare events should be recorded[20]. To be clear, the choice of approach to defining observable time is a critical decision that should be well documented for all observational studies. However, it is notable that increasing prominence of CDMs in health research has elevated this issue of defining observable time from a hidden, often undocumented study-specific parameter to an important underlying assumption applied to an entire database during the extract, transform, and load (ETL) process and thus applied across all study questions relying on those data.

This study is motivated by this shift in data processing standards. As defining observable time now requires explicit decisions and documentation during the ETL process, there is an urgent need to assess the implications of various methods. Here we examine multiple approaches to defining observable time across different types of databases with a focus on their impact on evidence generation by applying these methods to replicate a large study on background incidence rates of adverse events of special interest.

2 METHODS

2.1 Defining Eras of Observable Time

We use three different methods of defining eras of observable time based on observed health events, and one reference standard. Here we define observable time as the time during which any interaction with the healthcare system is captured.

Methods:

1. Persistence + surveillance approach
2. Min/Max
3. Age + gender approach

Reference Standard:

1. Enrollment time as a reference standard (enrollment-based databases only)

Health Event Identification

The creation of observation periods, based on patient interactions with the healthcare system requires the identification of periods of time when a person received care. To start, we identify unique dates with any healthcare event(s) per person. These events include visits, diagnoses, procedures, observations, device exposures, laboratory measurements, prescriptions written, drugs administered, and pharmacy drug dispensings. If a person was cared for in a tertiary setting for a length of time, then the start and end dates of the interaction are identified.

Often encounter-based databases are made up of EHR records that reflect interactions with a care provider in the office/hospital/outpatient center but do not extend to pharmacy interactions. In addition, some prescriptions are refilled automatically or by mail and do not necessarily represent a time when a person is actively receiving care. To address this fundamental difference and facilitate comparison between enrollment-based and encounter-based databases, we identify a second set of events (medical only) where those solely representing pharmacy drug dispensings are removed.

2.1.1 Persistence + Surveillance Approach

Building Observation Periods

We then create observation periods for each person using the identified dates with any healthcare event and dates with any medical event (a medical event is any healthcare event excluding pharmacy dispensings) by patient (hereafter referred to as "all events" and "medical events", respectively), by applying a combination of persistence and surveillance windows to the events. A persistence window is the maximum allowed number of days between event records that are then strung together to create an era of persistent observation. The surveillance window is the number of days added to the end of the persistent observation era as an additional period of surveillance prior to the end of the observation period. This is demonstrated in figure 1, which shows the difference in the observation periods between a surveillance window of 0 vs 90 days and persistence windows of 180 vs 365 days.

We apply 8 persistence windows of 180, 365, 548, 730, 1095, 1460, 1825, 2190 days combined with 5 surveillance windows of 0, 30, 90, 180, 270 days across each of the 11 databases for the two event types "all" and "medical". Overall, there are 80 combinations per database, 880 in total. Persistence and surveillance window values were chosen based on common patterns of healthcare utilization. For example, annual health checkups are required in Japan's JMDC database, and yearly physicals are common in U.S. clinical practice. These empirically informed intervals were selected to reflect realistic follow-up patterns and ensure consistency across databases, rather than through formal tuning or algorithmic selection.

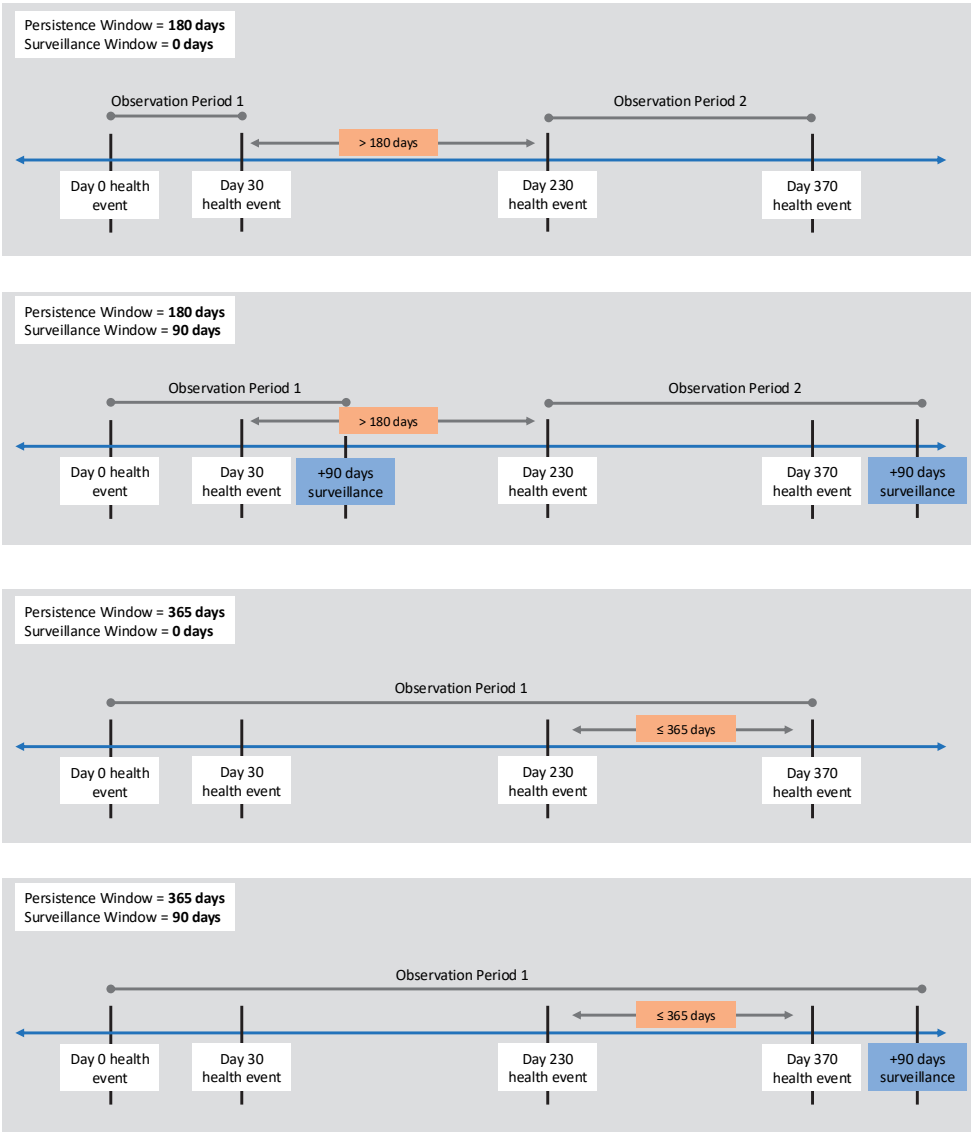


Figure 1: How persistence and surveillance windows are applied to health events to create observation periods

2.1.2 Age and Gender Approach

We created observation periods using dynamic persistence and surveillance windows tailored by age and gender, leveraging the distribution of time between healthcare events to account for demographic differences in healthcare utilization. For enrollment-based databases with near-complete capture of health events, we calculated the time in days between successive events for each individual, stratified by age and gender. The average time between events was computed per person and summarized to generate age- and

gender-specific distributions (mean, median, and 10th to 99th percentiles). This approach ensured equal weighting of individuals rather than events, preventing bias from chronically ill individuals with frequent visits. Observation periods were then inferred by applying the 99th percentile of the age- and gender-stratified distribution as the persistence window, while the 10th, 25th, 50th, 75th, and 90th percentiles were used as surveillance windows. This method was applied to both all healthcare events and medical events like the persistence + surveillance approach.

2.1.3 Min/Max

We also include the era defined as the earliest date a person was observed to the latest date a person was observed in the 5 databases without defined enrollment. This approach, otherwise known as "min/max", is the most commonly suggested approach for defining the observation period in encounter-based databases.

2.2 Enrollment Time as Reference Standard

We use the enrollment time as the reference standard for comparison in the 6 databases with defined enrollment. Due to the nature of enrollment-based databases it is the case that some percentage of the persons enrolled do not have any records in the database at all; these are known as non-care seeking individuals. These individuals still contribute person-time when calculating statistics like incidence rates but do not have any observed interactions or outcomes. Encounter-based databases only record patient interactions with the healthcare system and as such, do not have any non-care seeking individuals. To account for this fundamental difference in the makeup of the databases, we studied two more patient populations for the purposes of comparison in the study replication. One includes enrollment time for those patients with at least one health event (medical or pharmacy) and the other includes enrollment time for patients with at least one medical event.

2.3 Databases

In this study we use Merative™ Marketscan® Commercial Database (CCAE), Merative™ Marketscan® Multi-State Medicaid Database (MDCD), Merative™ Marketscan® Medicare Database (MDCR), Optum's Clinformatics® De-Identified Data Mart (Optum Extended), Optum® De-Identified Electronic Health Records (Optum EHR), JMDC, IQVIA® Adjudicated Health Plan Claims Data (Pharmetrics), IQVIA® Longitudinal Patient Database in Australia (Australia LPD), IQVIA® Disease Analyzer France (France DA), IQVIA ® Disease Analyzer Germany (German DA), and Premier. These databases all include feeds from multiple providers or are administrative claims that capture all billed healthcare encounters for covered persons and all databases used were standardized to the OMOP CDM v5.3 or higher. Full text descriptions of these databases can be found in Appendix 1. CCAE, MDCD, MDCR, Optum Extended, JMDC, and Pharmetrics are all enrollment-based while Optum EHR, Australia LPD, France DA, German DA, and Premier are encounter-based. Table 1 gives

high-level characteristics of each database used in the analysis including average age of patients, number of persons with no events (any and medical only as described in section 2.1), the provenance of the data and the time periods the databases span.

2.4 Study Replication

To evaluate the impact of the various observation time definitions, we replicate the study "Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study" by Li, et al. [21].

This study was originally designed to characterize the incidence rates of 15 different adverse events (AE) from 1 January 2017- 31 December 2019. Persons with at least one observation period that included the dates 1 January 2017, 1 January 2018, or 1 January 2019 with at least 365 days of observation prior to index were considered the target at risk population. The index date was defined as 1 January of each year. Of the original 15, five AEs were chosen for this replication: acute myocardial infarction, deep vein thrombosis, pulmonary embolism, anaphylaxis, and narcolepsy to represent rare, prevalent, acute, and less acute conditions. Links to the phenotype definitions can be found in Appendix 2.

A "clean window" was defined as the time period prior to the index date during which the AE could not be observed. If the AE was observed during that window, the participant was removed from the study cohort. The clean window for the replicated AEs was 365 days except for anaphylaxis, which was 30 days.

A person's time at risk was defined as 365 days after the index, censored on an event during the clean window, death, or the end of their observation period in the database. Incidence rates were calculated as the total number of events divided by the total time at risk per 100,000 person years.

We replicate the study across all 11 databases using the following definitions of the observation period and reference standard defined above:

- Persistence + Surveillance Method
  - 2 event types (all events, medical events)
  - 8 persistence windows
  - 5 surveillance windows
- Age + Gender Method
  - 1 persistence window
  - 5 surveillance windows
  - 2 event types (all events, medical events)
- Min/Max (encounter-based databases)
  - 2 event types (all events, medical events)

- Enrollment time as a reference standard (enrollment-based databases)
  - o For all patients
  - o Patients with any event
  - o Patients with any medical event

We then use the incidence rates (IRs) generated using enrollment time and compare them to the IRs generated using persistence + surveillance and the age + gender methods by calculating the mean square error (MSE). This allows us to determine the IRs that are the closest to the reference standard.

3 RESULTS

Tables 1 and 2 provide summary statistics for the enrollment-based and encounter-based databases, respectively. Across the 6 enrollment databases, the total number of persons ranged from 11,168,574 (MDCR) to 169,538,013 (CCAE). Between 94.58% (MDCR) and 82.89% (MDCD) had at least one event (medical or pharmacy) observed while 92.68% (JMDC) to 81.31% (Pharmetrics) had at least one *medical* event observed. In contrast, the total number of persons in the encounter databases ranged from 334,266,216 (Premier) to 1,258,464 (Australia LPD). Between 98.73% (France DA) and 100% (Premier) had at least one event observed (medical or pharmacy).

The median number of observation periods per person for all enrollment-based databases was 1 with median length in days from 545 in Optum Extended to 1,399 in JMDC. In the encounter-based databases, Premier had the lowest median days from earliest event to latest event at 20 days while Optum EHR had the highest at 1,170 days.

Australia LPD, France DA, and Germany DA have slightly less than 100% of persons with at least one event in the database with 98.84%, 98.73%, and 99.82%, respectively. Due to privacy-preserving measures implemented by the data vendor, death dates are removed from the dataset. In a small number of cases, death was the only recorded event for an individual. As a result, these individuals have no observable events in the data.

Table 1: Summary statistics for the enrollment-based databases; values are median (IQR) unless stated otherwise.

	Pharmetrics	JMDC	Optum Extended	CCAE	MDCD	MDCR
Persons						
Country	United States	Japan	United States	Japan	United States	United States
Total	159,649,500	16,572,535	97,963,738	169,538,013	35,775,653	11,168,574
Dates Available (YYYYMM)	201501 - 202307	200501- 202306	200005- 202309	200001- 202310	200601- 202212	200001- 202310
Median age at first observation period	33	31	34	31	18	70
Percent with one event (medical or pharmacy)	84.05%	92.81%	85.83%	84.71%	82.89%	94.58%
Percent with one medical event	81.31%	92.68%	83.02%	81.69%	81.81%	92.23%
Events per person (persons with any event)	16 (5-44)	43 (15-106)	21 (6-66)	25 (8-71)	44 (14-127)	107 (38-265)
Medical events per person (persons with any medical event)	12 (4-31)	43 (14-105)	14 (5-41)	19 (7-54)	39 (12-109)	72 (25-183)
Observation Periods*						
Percent of persons by number of observation periods						
1	88.79%	100.00%	78.51%	88.73%	73.25%	86.14%
2	9.56%	0.00%	16.48%	9.98%	17.35%	11.56%
3	1.36%	0.00%	3.88%	1.08%	5.63%	1.97%
4+	0.29%	0.00%	1.13%	0.21%	3.78%	0.33%
Percent of observation periods by length in years						
<=1	39.84%	18.31%	41.34%	41.68%	38.00%	30.07%
>1-2	21.36%	16.81%	20.99%	21.66%	19.29%	17.64%
>2-5	25.46%	29.29%	24.30%	23.45%	25.95%	29.76%
>5-10	13.34%	26.06%	10.14%	9.71%	14.25%	16.13%
>10	0.00%	9.52%	3.24%	3.50%	2.51%	6.40%
Events per observation period (persons with any event)	14 (4-39)	43 (15-106)	16 (5-50)	22 (7-63)	27 (8-80)	94 (34-227)
Medical events per observation period (persons with any medical event)	10 (3-28)	43 (14-105)	11 (4-31)	17 (6-48)	24 (8-69)	63 (22-158)

Table 1: Summary statistics for the enrollment-based databases; values are median (IQR) unless stated otherwise. (*continued*)

	Pharmetrics	JMDC	Optum Extended	CCAE	MDCD	MDCR
Total Time in Database**						
Length in days	600 (181-1416)	1237 (481-2182)	712 (206-1999)	579 (182-1402)	987 (276-2424)	998 (357-2204)

* Reflects the observation period based on enrollment time in a health plan

**Reflects the total time persons are observed in the database, from their first event to their last event (min/max)

Table 2: Summary statistics for the encounter-based databases; values are median (IQR) unless stated otherwise.

	Australia LPD	France DA	Germany DA	Optum EHR	Premier
Persons					
Country	Australia	France	Germany	United States	United States
Total	1,258,464	3,428,518	32,077,010	112,025,045	334,266,216
Dates Available	201805-202310	201809-202309	201310-202309	200701-202310	199912-202311
Median age at first observation period	37	36	45	37	42
Percent with one event (medical or pharmacy)	98.84%	98.73%	99.82%	100%	100%
Percent with one medical event	98.84%	98.73%	99.82%	99.97%	100%
Events per person (persons with any event)	4 (2-13)	4 (2-10)	4 (2-11)	12 (3-49)	4 (2-11)
Medical events per person (persons with any medical event)	4 (2-13)	4 (2-10)	4 (2-11)	11 (3-45)	4 (2-11)
Total Time In Database*					
Percent of persons by time in database in years					
<=1	51.42%	51.03%	55.35%	36.39%	99.91%
>1-2	17.71%	14.58%	7.56%	6.85%	0.07%
>2-5	28.55%	32.60%	15.51%	15.78%	0.02%
>5-10	2.32%	1.79%	19.32%	22.57%	0.00%
>10	0.00%	0.00%	2.38%	18.41%	0.00%

*Reflects the time persons are observed in the database, from their first event to their last event (min/max)

Figures 2 and 3 show how the incidence rate (IR) in CCAE and MDCR changed based on the persistence and surveillance window used as compared to that using enrollment time and the min/max approach. There is a clear decreasing trend in the IRs as both the persistence and surveillance windows increase, however, there is diminishing impact of modifying the persistence window once it exceeds 500 days. This holds across almost all databases and outcomes, with few exceptions as seen in Appendix 3.

Comparing the IRs generated using enrollment time and the min/max, in Pharmetrics, MDCD and CCAE they are very similar across the five AEs. In MDCR and Optum Extended the min/max IR consistently underestimates the enrollment IR while in JMDC the opposite is seen with the min/max IR consistently overestimating the enrollment IR.

In Optum EHR, Premier, and IQVIA Germany the min/max IR is low, corresponding to the estimates using 2,190 day persistence windows and 180 or 270 day surveillance windows. In contrast, the Min/Max estimates in IQVIA Australia LPD and IQVIA France fall around the midpoint in the range of IRs calculated using the persistence and surveillance approach.

For the enrollment-based databases, the IRs produced through the persistence + surveillance method were compared to the IRs produced using the enrollment time as the observation period by calculating the mean squared error (MSE). The MSE was then plotted as a heat map by outcome and event type with the min/max IR represented by a 9999-day persistence window and 0-day surveillance window. The darkest blue color on the graph corresponds with the lowest MSE. The MSE heat map for the Acute Myocardial Infarction outcome with the observation period defined using medical events only can be seen in figure 4; the additional heat map plots are available in appendix 4.

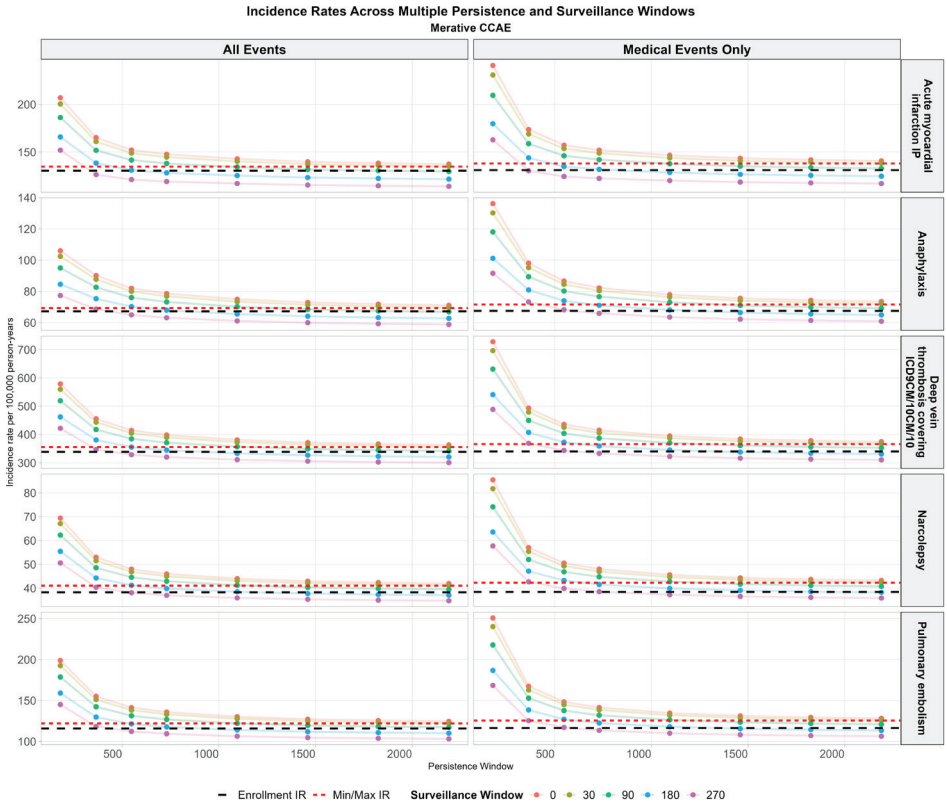


Figure 2: Incidence rates in Merative CCAE across multiple persistence and surveillance windows, five outcomes, and two event types

For all enrollment databases except JMDC, the IRs closest to those using enrollment time show an inverse relationship between persistence and surveillance, with the lowest persistence windows paired with the highest surveillance windows, and vice-versa. In Merative MDCR, the IRs produced using higher surveillance windows do not come as close to the enrollment IR as the IRs produced using 0- to 90-day surveillance windows though there is still a clear inverse relationship between persistence and surveillance.

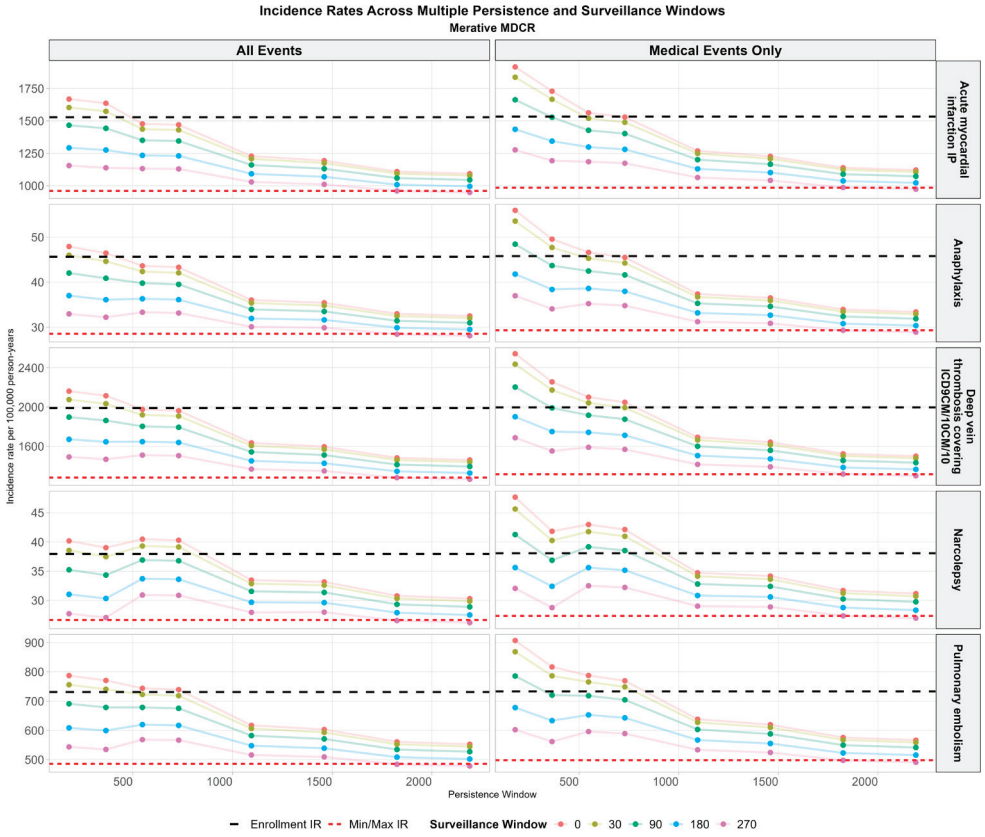


Figure 3: Incidence rates in Merative MDCR across multiple persistence and surveillance windows, five outcomes, and two event types

This same MSE approach was used to compare the enrollment IR to that using the age + gender approach, the results of which can be seen in figure 5 for the deep vein thrombosis outcome with the observation period defined using medical events only. Additional plots for the age + gender approach can be found in Appendix 5. The 99th percentile of average time between events was the only persistence window used and across all outcomes and databases, the surveillance window using the 90th percentile of average time between events produced IRs that came closest to the IRs produced using the enrollment time, especially when only medical events were used (except for Merative MDCR). A similar pattern in Merative MDCR observed in the persistence + surveillance approach was also observed in the age + gender approach, with the lower surveillance windows coming closer to approximating the IR using enrollment time. The age + gender approach did not perform as well in the rarer outcomes, specifically narcolepsy and anaphylaxis.

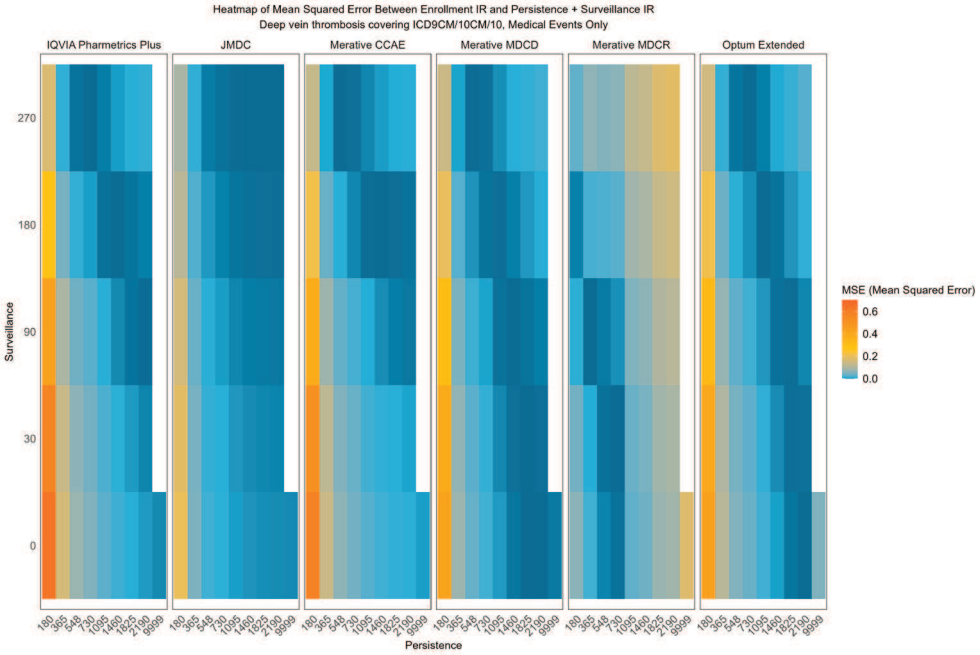


Figure 5: Heatmap of mean squared error between enrollment incidence rate and age + gender IR using medical events only for the outcome deep vein thrombosis

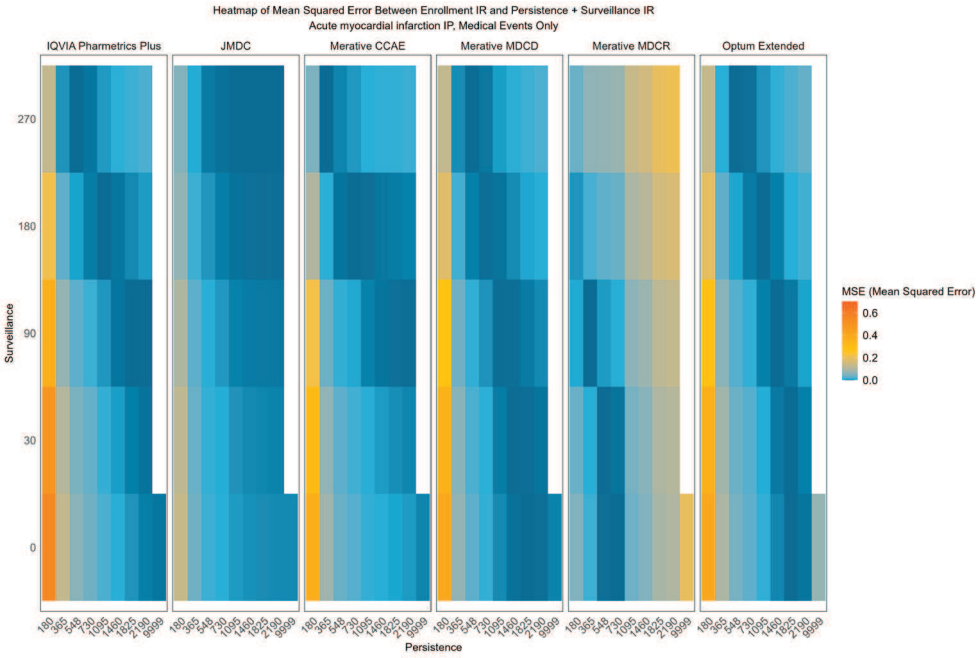


Figure 4: Heatmap of mean squared error between enrollment incidence rate and persistence + surveillance IR using medical events only for the outcome acute myocardial infarction.

4 DISCUSSION

When using real-world data for epidemiologic evidence generation the definition of observable time is critical and can lead to biased estimates if done incorrectly. Underlying almost every study relying on the secondary use of health data is a phenotype with some notion of "prior observation time" or a covariate balancing method that requires accurate look-back periods to produce balanced treatment and comparator arms [22–24]. Despite this, the calculation used to determine observable time per person is rarely discussed.

Without explicit enrollment or registration information inferring eras of observable time is especially difficult. The question is less about the completeness of the data, e.g. if the patient sought care with a different provider not captured by the data source, and more about the time span after an encounter when a patient is still considered observed. In other words, when a patient leaves one health visit, what is the span of time until their next expected visit? In real-world data It is possible to observe large spans of time with no health events and it is unknown whether these spans of time are truly devoid of events or if they are periods of incomplete data capture. Many encounter-based databases like EHRs use a first event to last event or min/max approach, but this assumes continuous observation during that time which may not be accurate.

To investigate how differing definitions of observable time in observational health databases impact evidence we conducted a large methodological experiment. By applying different persistence and surveillance windows (either static or based on age and gender) to string together health events into observation periods, we were able to test how incidence rates generated using these observation periods differed from each other as well as those generated using enrollment-time as a reference standard (enrollment-based databases only) and the min/max approach (encounter-based databases).

Our findings demonstrate a clear inverse relationship between observation window length and incidence rates: shorter windows resulted in higher incidence rates, while longer windows reduced them. This effect is intuitive, as increasing the denominator by extending observation time naturally lowers incidence rates. However, this phenomenon has not been empirically explored in detail before. Comparing the persistence + surveillance approach with enrollment time, we found that in U.S. claims databases (excluding Merative MDCR), the lowest persistence windows combined with the highest surveillance windows produced incidence rates closest to those obtained using enrollment time, and vice versa.

The Merative MDCR database showed differing trends when using the persistence + surveillance method and when comparing the IRs generated using this method to the enrollment IR. The requirement of a wash-out period of 365 days for many of the out-

comes studied caused the IRs to increase between the 180 and 365 persistence window estimates. This indicates that age plays a factor in how observation periods should be defined in encounter-based databases, but the age and gender model did not perform as well as expected. In future the age and gender model should be refined to better represent the healthcare utilization patterns of patients.

Furthermore, our comparison of the min/max approach with the persistence + surveillance method in encounter-based databases revealed that min/max-derived incidence rates varied substantially depending on the total duration of data availability. In databases with shorter time spans (e.g., IQVIA Australia LPD and IQVIA France, each with approximately five years of data), the min/max incidence rate estimates were typically in the mid-to-upper range of those generated using persistence + surveillance. In contrast, in databases with longer observational periods (e.g., IQVIA Germany, Optum EHR, and Premier), min/max estimates were closer to the lowest incidence rates produced using persistence + surveillance. This suggests that the longer duration of time available in an encounter-based database, the more susceptible the estimates are to error. Moreover, the min/max approach showed inconsistent results when compared with enrollment-based estimates, further questioning its reliability as a standardized method for defining observation periods.

While the above findings provide context for how to consider observable time at the database level, the replication of a study focused on incidence rates gives us insight into how these definitions influence measurement error. Often when estimating incidence rates, the focus is placed on accurately identifying cases. However, this study highlights another important source of measurement error: inaccuracies in defining the population at risk, particularly in databases without explicitly defined enrollment time. The assumptions made about how long an individual is "observable" within a dataset can substantially influence incidence estimates. This is especially problematic in encounter-based databases, where the length of available observation time varies greatly between patients. Our findings suggest that the longer the duration of data availability in an encounter-based database, the greater the susceptibility of incidence estimates to error, as prolonged observational periods may artificially reduce incidence rates by inflating the denominator.

Ultimately, our findings emphasize that the assumptions made about a database and its underlying population play a pivotal role in shaping incidence rate estimates. More research is needed to determine the most appropriate approach for defining observation periods, particularly in databases without explicit enrollment information. Developing standardized methodologies for inferring observable time, while considering the specific characteristics of different database types, will be essential to improving the reliability and comparability of incidence rate estimates in real-world evidence generation.

5 CONCLUSION

The observation period is a critical determinant of study validity, affecting both transparency and reproducibility in epidemiologic research. To improve data fitness-for-use, observation periods should be systematically assessed to ensure that the data capture process adequately supports answering clinical questions. Clearly defined observation period methodologies, tailored to specific database types, are essential to mitigating bias and enhancing the reliability of real-world evidence.

APPENDIX

Available online at https://pmc.ncbi.nlm.nih.gov/articles/PMC12361855/#sup1

DECLARATIONS

Author Contributions

CB, MJS, FD were involved in data collection. CB, MJS, PBR, MMC, PRR were involved in the study design, analysis, and interpretation of results. CB, MJS, MMC, PBR, PRR contributed to writing and all authors revised and approved the final draft.

Competing Interests

CB, MJS, FD, MMC, PBR, are employees of Johnson & Johnson and hold stock and stock options. PRR works for a department that receives/received unconditional research grants from Amgen, Chiesi, Johnson and Johnson, UCB Biopharma, the European Medicines Agency and the Innovative Medicines Initiative.

Patient Consent

Not required

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Data Availability

All data resulting from this work is available alongside the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

1. Mahmood SS, Levy D, Vasan RS, *et al.* The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective. *Lancet*. 2014;383:999–1008. doi: 10.1016/S0140-6736(13)61752-3
2. Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. *Curr Epidemiol Rep*. 2018;5:343–56. doi: 10.1007/s40471-018-0164-x
3. Optum Claims Data. https://www.optum.com/business/life-sciences/real-world-data/medical-pharmacy-claims-asset.html (accessed 23 January 2024)
4. Merative Marketscan Research Databases. https://www.merative.com/documents/brief/marketscan-explainer-general (accessed 23 January 2024)
5. Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol*. 2019;48:1740–1740g. doi: 10.1093/ije/dyz034
6. Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44:827–36. doi: 10.1093/ije/dyv098
7. Carey IM, Banchoff E, Nirmalananthan N, *et al.* Prevalence and incidence of neuromuscular conditions in the UK between 2000 and 2019: A retrospective study using primary care data. *PLoS One*. 2021;16:e0261983. doi: 10.1371/journal.pone.0261983
8. Weiskopf NG, Bakken S, Hripcsak G, *et al.* A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC)*. ;5. doi: 10.5334/egems.218
9. Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*. 2013;46:830–6. doi: 10.1016/j.jbi.2013.06.010
10. Lin KJ, Glynn RJ, Singer DE, *et al.* Out-of-system Care and Recording of Patient Characteristics Critical for Comparative Effectiveness Research. *Epidemiology*. 2018;29:356. doi: 10.1097/EDE.0000000000000794
11. Lin KJ, Singer DE, Glynn RJ, *et al.* Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data. *Clin Pharmacol Ther*. 2018;103:899–905. doi: 10.1002/cpt.861
12. Weber GM, Adams WG, Bernstam EV, *et al.* Biases introduced by filtering electronic health records for patients with "complete data." *Journal of the American Medical Informatics Association*. 2017;24:1134–41. doi: 10.1093/jamia/ocx071
13. The National Patient-Centered Clinical Research Network. Common Data Model (CDM) Specification, Version 6.1. 2023.
14. Observation Period Considerations for EHR Data. http://ohdsi.github.io/CommonDataModel/ehrObsPeriods.html (accessed 6 June 2023)
15. Eberly LA, Garg L, Yang L, *et al.* Racial/Ethnic and Socioeconomic Disparities in Management of Incident Paroxysmal Atrial Fibrillation. *JAMA Netw Open*. 2021;4:e210247. doi: 10.1001/jamanetworkopen.2021.0247
16. Spronk I, Korevaar JC, Poos R, *et al.* Calculating incidence rates and prevalence proportions: not as simple as it seems. *BMC Public Health*. 2019;19:512. doi: 10.1186/s12889-019-6820-3
17. Cheng C-Y. Risk of osteoporosis among individuals with varicose veins: a multi-institution cohort study. *Arch Osteoporos*. 2023;18:141. doi: 10.1007/s11657-023-01351-6
18. Tsai M-S, Lin M-H, Lee C-P, *et al.* Chang Gung Research Database: A multi-institutional database consisting of original medical records. *Biomed J*. 2017;40:263–9. doi: 10.1016/j.bj.2017.08.002

19. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574–8.
20. Observational Health Data Sciences and Informatics. OMOP Common Data Model (CDM). https://ohdsi.github.io/CommonDataModel/
21. Li X, Ostropolets A, Makadia R, *et al.* Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. *BMJ*. 2021;373:n1435. doi: 10.1136/bmj.n1435
22. Makadia R, Shoaibi A, Rao GA, *et al.* Evaluating the impact of alternative phenotype definitions on incidence rates across a global data network. *JAMIA Open*. 2023;6:ooad096. doi: 10.1093/jamiaopen/ooad096
23. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol*. 2018;47:2005–14. doi: 10.1093/ije/dyy120
24. Schneeweiss S, Rassen JA, Glynn RJ, *et al.* High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20:512–22. doi: 10.1097/EDE.0b013e3181a663cc

Data Quality

3

Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality

Clair Blacketer[1,2]
Frank J DeFalco[1]
Patrick B Ryan[1,3]
Peter R Rijnbeek[2]

[1]Observational Health Data Analytics, Janssen Research and Development, LLC, Titusville, NJ, USA
[2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[3]Department of Biomedical Informatics, Columbia University, New York, USA

Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021 Sep 18;28(10):2251-2257. doi: 10.1093/jamia/ocab132

ABSTRACT

Objective: Advances in standardization of observational healthcare data have enabled methodological breakthroughs, rapid global collaboration, and generation of real-world evidence to improve patient outcomes. Standardizations in data structure, such as use of Common Data Models (CDM), need to be coupled with standardized approaches for data quality assessment. To ensure confidence in real-world evidence generated from the analysis of real-world data, one must first have confidence in the data itself.

Materials and Methods: We describe the implementation of check types across a data quality framework of conformance, completeness, plausibility, with both verification and validation. We illustrate how data quality checks, paired with decision thresholds, can be configured to customize data quality reporting across a range of observational health data sources. We discuss how data quality reporting can become part of the overall real-world evidence generation and dissemination process to promote transparency and build confidence in the resulting output.

Results: The Data Quality Dashboard is an open-source R package that reports potential quality issues in an OMOP CDM instance through the systematic execution and summarization of over 3,300 configurable data quality checks.

Discussion: Transparently communicating how well CDM standardized databases adhere to a set of quality measures adds a crucial piece that is currently missing from observational research.

Conclusion: Assessing and improving the quality of our data will inherently improve the quality of the evidence we generate.

INTRODUCTION

As the amount of observational health data available to researchers continues to grow, regulatory agencies like the U.S. Food and Drug Administration (FDA)[1] and the European Medicines Agency (EMA)[2] have seen the value of using real-world evidence but there is still some concern and lack of trust in real-world data. Use of standardized data structures like the Observed Medical Outcomes Partnership Common Data Model (OMOP CDM)[3] have allayed some of these fears but standardization alone is not enough. Rigorous data quality assessments are needed to evaluate the quality of data with which evidence is generated.

The use of the OMOP CDM has lain the groundwork for impressive methodological and clinical breakthroughs in the areas of population-level effect estimation, patient-level prediction, and clinical characterization.[4–6] The recent *Large-Scale Evidence Generation and Evaluation across a Network of Databases* (LEGEND) study published in the Lancet is an exemplar of these ideas as the authors not only delivered relevant information about first-line antihypertensive drugs but also a novel approach to generating evidence using a systematic framework.[7] However, in order for regulatory agencies and clinicians to make decisions using such evidence, there needs to be trust not only in the methodologies employed but the underlying data itself.

The pitfalls of the secondary use of observational data to support research are well documented.[8–13] Typically these data are collected either for billing or diagnostic purposes and not with research endpoints in mind. Von Lucadou, et al. notes that information in the electronic health record may not be as granular as data captured during the course of a clinical trial and time stamps of clinical events should be examined prior to inferring temporal relationships.[9] Additionally, it is often the case that clinical ideas are captured in free-text fields, as described by Varela, et al.[11] It can be easy to overlook the multiple ways a white blood cell count is recorded, for example. One physician might use "WBC" while another uses "White Blood Cell" and yet another "White BC". These types of inconsistencies can lead to misclassification and measurement error.

Such issues are concerning but not unknown to clinical research networks (CRNs). The Sentinel Initiative,[14,15] the National Patient-Centered Clinical Research Network (PCORnet®),[16,17] and the Pediatric Learning Health System (PEDSnet),[18,19] among others, have all built processes and tools meant to identify data quality problems well in advance of any analytics that might be performed using the data. Historically, the Observational Health Data Sciences and Informatics (OHDSI) initiative promoted a tool known as the Automated Characterization of Health Information at Large-scale Longitudinal Evidence System (ACHILLES) to assess the quality of data in the OMOP CDM format.[20] ACHILLES is primarily used for database characterization. It computes a set of aggregate summary

statistics such as gender and age stratifications of persons included, average follow-up time, and distribution of diagnosis codes, among others. These statistics are then assessed for quality by running the ACHILLES Heel rules against the aggregated summaries rather than on the database itself. These rules include looking for patients with an age less than zero and prescription dispensing records with implausible drug quantities.

A comparison of the data quality assessment (DQA) programs across six different CRNs in 2017 revealed that OHDSI had the fewest data quality checks in place (172) while the other networks ranged from 875 up to 3,434 checks.[21] The reason for this difference is that OHDSI as an open collaborative has traditionally left DQA to the individual data owners. Lead protocol investigators are responsible for making a "fitness-for-use" decision and independently determine if a dataset is suitable to answer a clinical question. As OHDSI continues to move in the direction of large-scale network research[22–24] a more robust data quality tool is needed to ensure that a participant's data comply with community defined standards.

It is our goal to address the need for better data quality processes with the development of the Data Quality Dashboard (DQD). In this paper we describe the methods used to design the tool through community engagement. Next, we describe the inner workings of the DQD and run it against a US claims database as a proof-of-concept to show utility in practice. Finally, we discuss future enhancements and the potential for this tool to change the way observational data is utilized within OHDSI and beyond.

BACKGROUND AND SIGNIFICANCE

To take advantage of the tools and methodologies[4,25] available to the OHDSI community, collaborators must first convert their data to the OMOP CDM. Data owners, clinicians, and OMOP experts all come together to standardize a source database by putting it in the structure of the model and applying agreed upon conventions through a process known as extract, transform, load (ETL). This is the expected function of any standard data model but where OMOP differs is that not only the structure but also the content of the data is standardized, harmonizing on the SNOMED vocabulary[26] for conditions and RxNorm vocabulary[27] for drugs, for example. It is this semantic standardization that facilitates international adoption and fosters rapid collaboration.

Once a database is converted to the CDM ideally it should be assessed for quality prior to using it for generating evidence. Some collaborators, like PEDSNet, designed their own data quality tools to keep track of metrics of interest to them.[19] For many collaborators, the ACHILLES Heel report served this function up until now. The tool runs a set of checks against a CDM instance and reports them back to the user as an available option in the

Achilles characterization package[1]. The Heel report gives an overview of potential data quality issues, but it does not allow the user to change how a pass or fail for a given check was determined, does not share the Structured Query Language (SQL) query that was run to produce the result, nor does give the option to capture any metadata about why a perceived quality issue might be occurring. There is also no process for extending the tool either by adding features or new data quality checks.

We therefore developed the DQD as a stand-alone R package to improve upon prior work and to fill in the gaps left by the Achilles Heel report. By focusing our attention to assessing data after the conversion to the OMOP CDM, the standardized structure and content directly enabled the creation of a standardized quality control framework with the DQD at the center (figure 1).

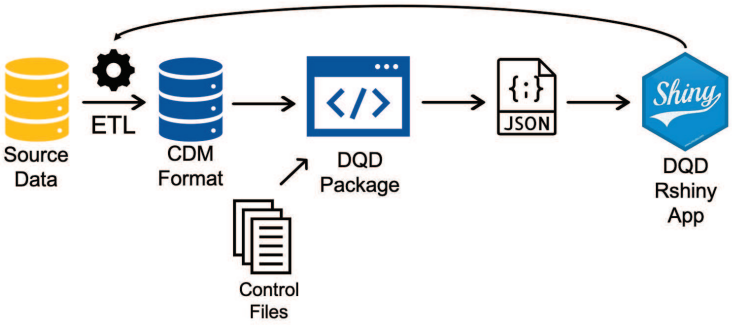


Figure 1: The standard quality control framework

MATERIALS AND METHODS

To lead this initiative a few interested OHDSI collaborators formed a committee to identify high-level data quality issues important to the community. We aligned with the framework described by Kahn, et. al.[28] as a way to organize our approach. Kahn and his colleagues identified the categories of conformance, completeness, and plausibility into which most, if not all, data quality checks can be grouped.

Conformance checks measure how well a database conforms to specified formats and relational constraints. The committee applied this idea to the OMOP CDM, describing issues such as whether a CDM instance contains all required fields, if fields that are defined as primary keys contain unique values, and if a foreign key value is present in its corresponding primary key field.

1 https://github.com/ohdsi/achilles

Completeness checks look at the frequency of values in a given dataset without examination of the values themselves. In terms of the CDM the idea of completeness can likewise be used to understand the quality of vocabulary mapping. The committee identified this as an area of importance for an OMOP data quality solution, including checks to evaluate the proportion of source values (ICD10CM, CPT4, Read, etc.) that were not mapped to standard concepts.

Plausibility checks are meant to gauge the believability of values in a dataset. This can take many forms like making sure a person's health care encounters all occur on or after their birthdate or looking to see that no one has a weight of zero kilograms recorded.

After deciding on the data quality checks to implement in a new data quality tool the committee turned their sights to the features they would like included. These were chosen based on discussions with key stakeholders from both academia and industry about their needs in this space. Considering the myriad of infrastructure constraints present in the community it should be a stand-alone application that any user can run out-of-the box. It should be easily scalable to allow inclusion of additional checks over time. It should also be flexible to allow adjustment of the checks and failure thresholds based on apriori knowledge of a database. The results of the data quality assessment should be easily shareable, in some form or fashion that is not a burden to the data owner.

With these requirements in mind, we devised a framework that would allow us to use the structure of the data model to our advantage. Since we already know the schema of every database that will run this tool, we didn't have to focus on how to write and execute individual data quality checks; instead, we were able to abstract a layer and define data quality *ideas*. For example, what if it is important to assess the number of persons in the Person table that don't have a record in the Visit_Occurrence table? That check can be assessed using a simple SQL statement against an OMOP CDM instance:

```
SELECT COUNT (DISTINCT P.person_id)
    FROM PERSON as P
    LEFT JOIN VISIT_OCCURRENCE as V
    ON P.person_id = V.person_id
    WHERE V.person_id IS NULL
```

As a data quality idea we are simply evaluating the degree to which persons in the Person table are represented in a fact table (Visit_Occurrence in this example). In the OMOP CDM the field Person.person_id is a primary key with corresponding foreign keys in all clinical fact tables. Using that constraint as our guide, the abstraction of this data quality check to a data quality idea results in a SQL statement like this:

```
SELECT COUNT (DISTINCT P.person_id)
    FROM PERSON as P
    LEFT JOIN @cdmTable
    ON P.person_id = @cdmTable.person_id
    WHERE @cdmTable.person_id IS NULL
```


Where the object @cdmTable represents the universe of clinical fact tables that have a foreign key to the Person table. Each of them can be rotated into the SQL statement to take the place of the parameter. If there are 15 of such tables then the above SQL would automatically generate 15 data quality checks from one data quality idea.

Using the process described above as our foundation we developed the Data Quality Dashboard (DQD) R package that systematically runs and evaluates data quality checks based on the structure of the Common Data Model and pre-specified failure thresholds.

To test it, the DQD was run on the IBM Marketscan® Multi-State Medicaid (MDCD) database. This database contains adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses and procedures, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. The major data elements contained within this database are outpatient pharmacy dispensing claims, inpatient, and outpatient medical claims. The data does not contain laboratory results.

RESULTS

In total, 20 high-level data quality ideas were identified: 5 evaluating completeness, 8 evaluating conformance, and 7 evaluating plausibility, the complete list of which is included in appendix 1. The example given above to assess the extent to which persons in the Person table are represented in the clinical fact tables became the data quality idea, or check type, *measurePersonCompleteness*. In terms of OMOP, completeness means not only missingness but vocabulary mapping completeness. The results of two check types, *standardConceptRecordCompleteness* (SCRC) and *sourceValueCompleteness* (SVC), work together to show how well the diagnostic, procedural, and drug codes, etc. (source values) in a database have been mapped to the standard terminology as defined by OMOP. SCRC counts the number of records for a given table that have been mapped to a Concept_Id of zero where a Standard Concept is expected. If applied to the Condition_Occurrence table, for example, it would count the number of records with a Condition_Concept_Id of zero. In a situation where a source value in a database cannot be mapped to a Standard Concept, a zero is used to denote "No matching concept". Therefore, if a large number of

records in a table have a Standard Concept of zero, those records cannot be used in an analysis as standardized analytics rely on standardized vocabularies.

The SVC check type, on the other hand, quantifies the number of distinct source values in a database that have been mapped to zero. Using the example from earlier, the values "WBC", "White Blood Cell", and "White BC" are all ways in which the clinical idea of a white blood cell count measurement might be represented. These values are all free text and as such do not have an automatic mapping to a Standard Concept. Thus, records with these values would be given a Standard Concept Id of zero during the ETL conversion process. If database A has 5,000 records in the resulting Measurement table with these source values, then the SCRC check type applied to the Measurement_Concept_Id field in the Measurement table would return 5,000 records mapped to zero while the SVC check type applied to the same field would return 3 distinct values mapped to zero. This can be interpreted to mean there are 5,000 records representing 3 distinct values that are missing a mapping to a Standard Concept in the Measurement_Concept_Id field of database A. The relationship between these two checks will signal different vocabulary or ETL changes to be made. A high SCRC and high SVC might indicate that the database has a proprietary coding system not represented in the OMOP vocabulary. In such case either the coding system is added to the vocabulary or the source codes are manually mapped to standard concepts. A high SCRC and low SVC is often due to a small number of catch-all values like "unknown" or "UNK". These do not have any real meaningful analytic use so the records are either ignored (by increasing the failure threshold) or they are removed, in which case they would no longer be visible to the DQD. A low SCRC and high SVC usually means that the source values representing the highest number of records were mapped to standard concepts and the rest were mapped to zero. This is an accepted ETL practice as the records mapped to zero are retained for later use if necessary. A low SCRC and low SVC means that a large number of source codes representing a large number of records were mapped to standard concepts. This is the ideal scenario for an ETL though it is important to monitor these values over time in the event the source coding practices change.

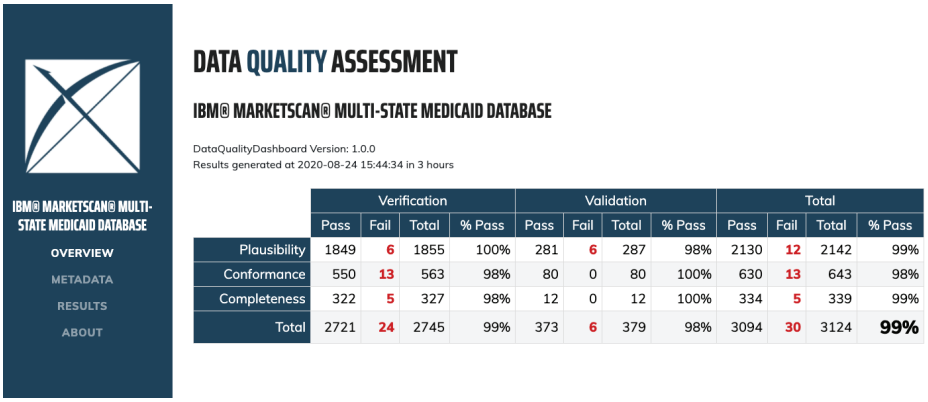
The *plausibleTemporalAfter* check type evaluates the temporal relationships between date values. Applied to the Visit_Occurrence table this check quantifies the number of visits with a visit end date prior to the visit start date. If such visits are found, the ETL must make a choice how to handle them. The most common practice is to choose either the start or end date as listed in the source data and use that value for both fields (Visit_Start_Date, Visit_End_Date) in the standardized dataset. This allows for the information from the visit to be retained while eliminating the temporal inconsistency.

When the *plausibleTemporalAfter* check type is applied to the Condition_Occurrence table it becomes a way to measure the number of condition records that occur prior to a person's birth. There can be many reasons for records to be written with incorrect dates but usually these are patient history records where no date was given by the patient and, instead, a default date is assigned by the electronic health record. As it impossible to discern the correct date for these records the typical recourse is to remove them.

These check types were all written as parameterized SQL statements. To resolve and then run these SQL statements the DQD reads a set of included control files that detail each table and field in the CDM, their constraints, and their relationships to one another. These files also indicate which data quality check types should be run on which fields and what thresholds should be applied to the results to determine a pass or fail, all of which can be edited by the user. The tool then takes these files and swaps out the parameters in the SQL statements for the values indicated in the control files using the SqlRender R package.[29]

With this approach the 20 checks types are resolved to over 3,300 individual quality checks: 396 evaluating completeness, 779 evaluating conformance, and 2126 evaluating plausibility. The DQD then compiles all results into a text- based file as the default output, though the user can also specify the option to write the results back to a table in the database. The resulting file (or table) contains all information produced by the tool and is read by the package to render an interactive Rshiny[30] user interface (figure 1).

The first screen that greets the user is the overview tab, shown in figure 2 displaying the result of running the DQD on the MDCD database. The tables Note, Note_Nlp, and Specimen are not populated for MDCD, resulting in a lower total number of checks at 3,124 instead of 3,300. The total run time of the DQD was 7 hours for this database sized around 700GB in a cloud-based parallel environment.



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1849	6	1855	100%	281	6	287	98%	2130	12	2142	99%
Conformance	550	13	563	98%	80	0	80	100%	630	13	643	98%
Completeness	322	5	327	98%	12	0	12	100%	334	5	339	99%
Total	2721	24	2745	99%	373	6	379	98%	3094	30	3124	**99%**

Figure 2: The overview tab of the DQD showing data quality check pass percentages by category in MDCD.

The overview tab gives a high-level summary of the total number of passes and failures by Kahn category. In figure 2, the database examined has 13 conformance failures, or instances where it does not conform to the specifications of the OMOP CDM. It has 5 completeness failures related to potentially missing data and 12 plausibility failures which could be a myriad of issues including incorrect dates or implausible measurement values. To explore each of these failures the results tab (figure 3) shows one line per check run. Across the top is the option to filter by pass/fail status, CDM table, and Kahn context. The plus sign on each line expands to show the exact SQL query that was run to achieve that result, allowing the user to pinpoint the identified failing records in their dataset. A publicly available instance of these data quality results can be found at https://data.ohdsi.org/DataQualityDashboardMDCD/ and the text file with the full result set is available in supplementary materials.

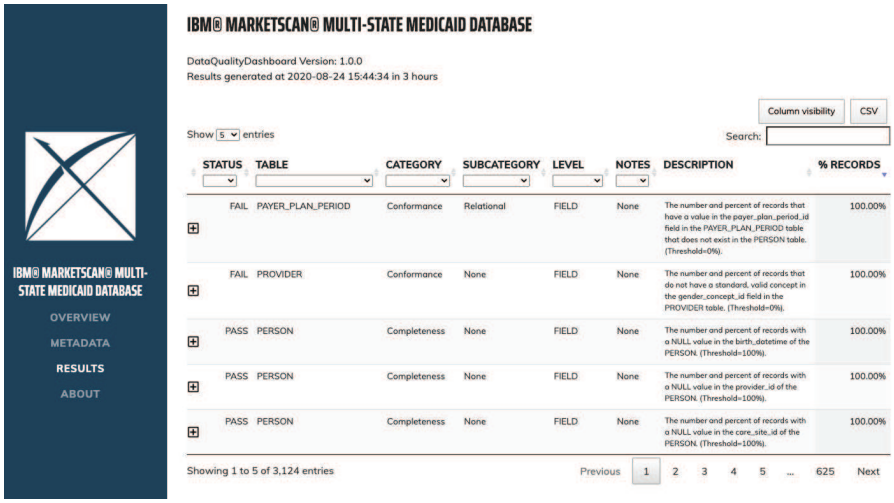


Figure 3: DQD results page with all data quality checks run and their outcomes.

One of the major findings from this exercise showed a high SCRC (19.20%) and high SVC (23.34%) for the Procedure_Concept_Id field of the Procedure_Occurrence table. Further investigation found the problem was due to a previously unaccounted for source vocabulary. Unlike most commercial health insurance plans, state Medicaid coverage usually includes routine dental care. These procedures are coded using Current Dental Terminology (CDT). This had not been documented in the ETL so all dental records were being mapped to a Procedure_Concept_Id of zero. To fix this issue the CDT vocabulary was added to the ETL and the ETL was re-run.

In this case a full batch-job was completed in which the entirety of the ETL and DQD were re-run and examined. In absence of a fix, ETLs are most often run as the source data are updated. Increasing numbers of data owners have moved to an incremental model where only the newly acquired data are run through the ETL process as this method is both time- and cost-effective. While research is ongoing in this area the current best practice is to break the DQD into two steps. As data is initially mapped to the OMOP CDM and the ETL is set up, only conformance and completeness checks are run to ensure the newly minted CDM database conforms to the standards set by the OHDSI community. Once those checks are satisfied the plausibility checks are run initially and each time an incremental load is processed. The underlying assumption is that once conformance and completeness is satisfied, the structure of the database should not change as long as the version of the CDM and vocabulary does not change. Plausibility should be assessed with each load to confirm that the new data complies with the gender, temporal, and measurement value quality measures.

The DQD is available on GitHub[31] as a fully executable R[32] package, supporting OMOP CDM versions 5.2.2 and 5.3.1. It works with multiple database management systems including PostgreSQL, Microsoft SQL Server, and Redshift. All documentation can be found at https://ohdsi.github.io/DataQualityDashboard/.

DISCUSSION

As use of observational health data continues to increase, two of the largest regulatory agencies in the world have established plans for how these data can be used to support data-driven regulatory decision-making. The U.S. Food and Drug Administration (FDA) published the Framework for FDA's Real-World Evidence Program in 2018 which details how and in what capacity real world evidence generated from real world data might be evaluated for its potential use to support approvals of new drug indications or to satisfy post-marketing safety studies.[1] Similarly, the Heads of Medicines Agency (HMA) and the European Medicines Agency (EMA) in 2017 initiated a joint Big Data Task Force. A summary report from 2019 describes a strategy to understand observational data such

that they might be ready to make use of it in a regulatory capacity.[2] These reports are robust, especially when discussing how real-world evidence supporting drug safety and effectiveness research should be conducted. Both groups agree that data quality is important and should be considered when determining whether a set of data is suitable to answer specific questions. This is especially timely given the COVID-19 pandemic and rush to publish any information that may further understanding of the natural history and clinical treatment of the disease.[33]

The Data Quality Dashboard is well-poised to answer the call for higher quality data. It is at the center of the standard quality control framework (figure 1), providing critical insights to data owners that then feed back into the ETL process, ultimately resulting in research-ready data. There is evidence of broad adoption of the DQD at the network level as the European Health Data and Evidence Network (EHDEN)[34] and the National COVID Cohort Collaborative (N3C)[35] both leverage the framework to ensure that participating databases pass critical quality control measures. As these networks conduct research and learn from their data, any quality issues identified during analysis will be incorporated back into the tool. Continuous iteration on the set of data quality checks in coordination with research creates a living system that improves as we advance our understanding. The DQD will also continue to add features and expand the check types it covers. Initial roadmaps include enabling it to be run on a cohort rather than the entire database, extending the check types to include the OMOP vocabulary, incorporating prevalence measures, and evaluating temporal stability.

Other CRNs have built tools and processes for evaluating and managing the quality of data at the network level. The PEDSNet DQ Workflow [36] utilizes a similar software architecture in that it relies heavily on R, is designed to be applied to databases on the OMOP CDM standard (modified slightly for pediatric use) and can be run on multiple different database management systems (PostgreSQL, Oracle, MySQL, SQLite, or SQL Server). PEDSnet is supported by the Patient Centered Outcomes Research Institute (PCORI) and brings together data from various hospitals into one large, pediatric database. This structure is reflected in the DQ Workflow as the coordinating center has more control over how data quality issues are reported and tracked across the network. Sites run the quality assessment, the results of which are then sent to the coordinating center. Results are reviewed by a data scientist who discusses any issues and corrections needed with the site. The Sentinel Initiative [14,15] has a similar system where Data Partners run a data quality study package, results are sent to the Sentinel Operations Center, and the Operations Center reviews and recommends changes to the ETL. [37] Sentinel data quality checks are divided into different levels and each Data Partner must pass all level 1 checks before moving on to level 2, etc. PCORNet has a five-step data curation process that also involves review of the data quality results by a Coordinating Center. This review is done in cycles

whereby network partners are expected to correct any model conformance issues before running completeness checks, for example.

OHDSI, in contrast, is a distributed data network with no central coordination of data. Each member is responsible for their own data and ETL processes. This is a decided choice that allows for faster, broader reaching collaboration. The OMOP CDM could of course be used in a distributed network but sharing protocols and analytic code rather than data eliminates the need to take data beyond the firewall of the data holder. Such a system allows far more sites around the world to contribute as there is no longer an issue of data governance. Therefore, to ensure the quality of the OHDSI network, the DQD has to both assess and communicate the quality of a database while taking differences at the source into account. This led to a tool built using a dynamic framework that expands 20 check types into over 3,300 individual checks. Data owners have full control over which checks are run and how they are assessed for failure while the resulting report details every single check, pass or fail, in an easy-to-share way. It is our vision that this becomes the way data quality is shared among networks, with reviewers, and with regulators. The most recent papers published using the PEDSnet, PCORNet, and Sentinel networks [38–40] either do not mention how the data used were assessed for quality or they touch on it very briefly. Instead, with a tool like the DQD, a file can be shared as additional material to a publication detailing all data quality checks that were run on the database in which the study was executed. This level of transparency is unprecedented in current literature but is necessary in an era where we are asked to trust evidence generated by the scientific community with little to no insight into the data used.

CONCLUSION

The Data Quality Dashboard has established a new way to garner trust in real-world data. Transparently communicating how well CDM standardized databases adhere to a set of quality measures adds a crucial piece that is currently missing from observational research. Assessing and improving the quality of our data will inherently improve the quality of the evidence we generate.

SUPPLEMENTARY MATERIAL

Available online at https://pmc.ncbi.nlm.nih.gov/articles/PMC8449628/#sup1

DECLARATIONS

Funding

This study was sponsored by Janssen Research & Development, LLC. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Conflicts of interest/Competing interests

CB, FJD, and PBR are Janssen employees and own stock or stock options. PRR receives an unconditional grant from Janssen Research & Development, and funding through the Innovative Medicines Initiative.

Ethics approval

The use of IBM MarketScan® Multi-State Medicaid Database was reviewed by the New England Institutional Review Board and determined to be exempt from broad Institutional Review Board approval, as this research project did not involve human patient research.

Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The Data Quality Dashboard R package is available at https://github.com/OHDSI/DataQualityDashboard/. A publicly available instance of the data quality results for IBM MarketScan® Multi-State Medicaid can be found at https://data.ohdsi.org/DataQualityDashboardMDCD/.

Author Contributions

Conception and design - all authors, collection and assembly of data - all authors but primarily CB, data analysis and interpretation - all authors but primarily CB manuscript writing - all authors, and final approval of manuscript - all authors.

REFERENCES

1. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018.
2. Heads of Medicines Agency, European Medicines Agency. HMA-EMA Joint Big Data Taskforce Summary Report. 2019;:48.
3. OMOP Common Data Model. http://ohdsi.github.io/CommonDataModel/ (accessed 14 Jan 2021).
4. Reps JM, Schuemie MJ, Suchard MA, *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc JAMIA* 2018;25:969–75. doi:10.1093/jamia/ocy032
5. Ryan PB, Schuemie MJ, Gruber S, *et al.* Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Saf* 2013;36 Suppl 1:S59-72. doi:10.1007/s40264-013-0099-6
6. Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016;113:7329–36. doi:10.1073/pnas.1510502113
7. Suchard MA, Schuemie MJ, Krumholz HM, *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet Lond Engl* 2019;394:1816–26. doi:10.1016/S0140-6736(19)32317-7
8. Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit Transl Bioinforma* 2010;2010:1–5.
9. von Lucadou M, Ganslandt T, Prokosch H-U, *et al.* Feasibility analysis of conducting observational studies with the electronic health record. *BMC Med Inform Decis Mak* 2019;19:202. doi:10.1186/s12911-019-0939-0
10. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf Med* 2009;48:38–44.
11. Otero Varela L, Wiebe N, Niven DJ, *et al.* Evaluation of interventions to improve electronic health record documentation within the inpatient setting: a protocol for a systematic review. *Syst Rev* 2019;**8**:54. doi:10.1186/s13643-019-0971-2
12. Alla F, Rosilio M, Funck-Brentano C, *et al.* How can the quality of medical data in pharmacovigilance, pharmacoepidemiology and clinical studies be guaranteed? *Therapie* 2013;68:209–23. doi:10.2515/therapie/2013040
13. Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Ann Intern Med* 2020;172:S79–84. doi:10.7326/M19-0873
14. Behrman RE, Benner JS, Brown JS, *et al.* Developing the Sentinel System--a national resource for evidence development. *N Engl J Med* 2011;364:498–9. doi:10.1056/NEJMp1014427
15. Ball R, Robb M, Anderson SA, *et al.* The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther* 2016;99:265–8. doi:10.1002/cpt.320
16. Collins FS, Hudson KL, Briggs JP, *et al.* PCORnet: turning a dream into reality. *J Am Med Inform Assoc JAMIA* 2014;21:576–7. doi:10.1136/amiajnl-2014-002864
17. Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *eGEMs*;6. doi:10.5334/egems.199
18. Forrest CB, Margolis PA, Bailey LC, *et al.* PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc JAMIA* 2014;21:602–6. doi:10.1136/amiajnl-2014-002743
19. Utidjian LH. Identifying and Understanding Data Quality Issues in a Pediatric Distributed Research Network. American Academy of Pediatrics 2015. https://aap.confex.com/aap/2015/webprogrampress/Paper30131.html (accessed 13 Jan 2021).

20. Huser V, DeFalco FJ, Schuemie M, *et al.* Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS Wash DC* 2016;4:1239. doi:10.13063/2327-9214.1239
21. Callahan TJ, Bauck AE, Bertoch D, *et al.* A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *EGEMS Wash DC* 2017;5:8. doi:10.5334/egems.223
22. Morales DR, Conover MM, You SC, *et al.* Renin-angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis. *Lancet Digit Health* Published Online First: 17 December 2020. doi:10.1016/S2589-7500(20)30289-2
23. Schuemie MJ, Ryan PB, Pratt N, *et al.* Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study. *J Am Med Inform Assoc JAMIA* 2020;27:1268–77. doi:10.1093/jamia/ocaa124
24. Golozar A, Lai LY, Sena AG, *et al.* Baseline phenotype and 30-day outcomes of people tested for COVID-19: an international network cohort including >3.32 million people tested with real-time PCR and >219,000 tested positive for SARS-CoV-2 in South Korea, Spain and the United States. *MedRxiv Prepr Serv Health Sci* Published Online First: 27 October 2020. doi:10.1101/2020.10.25.20218875
25. Schuemie M, Suchard M, Ryan P. *CohortMethod: New-user cohort method with large scale propensity and outcome models*. 2020.
26. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html (accessed 4 Mar 2021).
27. UMLS Metathesaurus - RXNORM (RXNORM) - Synopsis. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/RXNORM/index.html (accessed 4 Mar 2021).
28. Kahn MG, Callahan TJ, Barnard J, *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS Wash DC* 2016;4:1244. doi:10.13063/2327-9214.1244
29. Schuemie M, Suchard M. *SqlRender: Rendering Parameterized SQL and Translation to Dialects*. 2020.
30. Chang W, Cheng J, Allaire JJ, *et al. shiny: Web Application Framework for R*. 2020. https://CRAN.R-project.org/package=shiny
31. Blacketer C, DeFalco F, Londhe A, *et al.* DataQualityDashboard. https://ohdsi.github.io/DataQualityDashboard/ (accessed 21 Jan 2021).
32. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: : R Foundation for Statistical Computing 2013. http://www.R-project.org/
33. Brainard J. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? Sci. AAAS. 2020.https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat (accessed 21 Jan 2021).
34. European Health Data Evidence Network. ehden.eu. https://www.ehden.eu/ (accessed 12 Mar 2021).
35. Haendel MA, Consortium the N, Chute CG, *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* doi:10.1093/jamia/ocaa196
36. Khare R, Utidjian LH, Razzaghi H, *et al.* Design and Refinement of a Data Quality Assessment Workflow for a Large Pediatric Research Network. *eGEMs*;**7**. doi:10.5334/egems.294
37. How Sentinel Gets Its Data. Sentin. Initiat. 2019.https://www.sentinelinitiative.org/about/how-sentinel-gets-its-data (accessed 9 Feb 2021).
38. Bailey LC, Razzaghi H, Burrows EK, *et al.* Assessment of 135 794 Pediatric Patients Tested for Severe Acute Respiratory Syndrome Coronavirus 2 Across the United States. *JAMA Pediatr* 2021;175:176–84. doi:10.1001/jamapediatrics.2020.5052

39. Bachmann KN, Roumie CL, Wiese AD, *et al.* Diabetes medication regimens and patient clinical characteristics in the national patient-centered clinical research network, PCORnet. *Pharmacol Res Perspect* 2020;**8**:e00637. doi:10.1002/prp2.637
40. Eworuke E, Haug N, Bradley M, *et al.* Risk of Non-Melanoma Skin Cancer in Association with Use of Hydrochlorothiazide-Containing Products in the United States. *JNCI Cancer Spectr* Published Online First: 4 February 2021. doi:10.1093/jncics/pkab009

4

Using the Data Quality Dashboard to Improve the EHDEN Network

Clair Blacketer [1,2,3*,†]
Erica A. Voss [1,2,3,*,†]
Frank DeFalco [1,3]
Nigel Hughes [1,3]
Martijn J. Schuemie [1,3]
Maxim Moinat [2,3,4]
Peter R. Rijnbeek [2,3]

1 Janssen Pharmaceutical Research and Development LLC, Titusville, NJ 08560, USA
2 Department of Medical Informatics, Erasmus University Medical Center, 3015 GD Rotterdam, The Netherlands
3 OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York 10027, NY, USA
4 The Hyve, 3511 MJ Utrecht, The Netherlands
* Authors to whom correspondence should be addressed.
† Co-first authors.

Blacketer C, Voss EA, DeFalco F, Hughes N, Schuemie MJ, Moinat M, Rijnbeek PR. Using the Data Quality Dashboard to Improve the EHDEN Network. Applied Sciences. 2021; 11(24):11920. https://doi.org/10.3390/app112411920

ABSTRACT

Federated networks of observational health databases have the potential to be a rich resource to inform clinical practice and regulatory decision making. However, the lack of standard data quality processes makes it difficult to know if these data are research ready. The EHDEN COVID-19 Rapid Collaboration Call presented the opportunity to assess how the newly developed open-source tool Data Quality Dashboard (DQD) informs the quality of data in a federated network. Fifteen Data Partners (DPs) from 10 different countries worked with the EHDEN taskforce to map their data to the OMOP CDM. Throughout the process at least two DQD results were collected and compared for each DP. All DPs showed an improvement in their data quality between the first and last run of the DQD. The DQD excelled at helping DPs identify and fix conformance issues but showed less of an impact on completeness and plausibility checks. This is the first study to apply the DQD on multiple, disparate databases across a network. While study-specific checks should still be run, we recommend that all data holders converting their data to the OMOP CDM use the DQD as it ensures conformance to the model specifications and that a database meets a baseline level of completeness and plausibility for use in research.

Keywords: data quality; OMOP CDM; EHDEN; healthcare data; real world data; RWD

1. INTRODUCTION

Over the past 30 years, more observational health data have become available for use in research due to the digitization of health records and the comprehensive nature of administrative claims [1,2,3]. The potential of this data is well known; the non-invasive, passive method of data collection bypasses the ethical concerns of human subjects research and the speed with which it is collected foreshadows a future of near-real time evidence generation [4,5]. Professional organizations such as the American Thoracic Society have publicly announced their intention to use observational studies to inform clinical practice guidelines [6]. Similarly, both the United States Food and Drug Administration and the European Medicines Agency have begun to rely more heavily on real-world evidence (RWE) to support critical drug safety decisions [7,8].

To maximize the research capability of such large-scale observational health data, the Innovative Medicines Initiative made funding available to develop a network of health data sources to support outcomes-focused healthcare delivery in Europe [9]. This project is known as the European Health Data and Evidence Network (EHDEN) [10]. As of October 2021, there are 98 databases in the network representing a total of 23 countries and about 450 million patient records. With the rise of COVID-19, this need for a European-wide federated network of observational health data was thrown into sharp relief as health officials scrambled to understand the natural history of the disease [11]. To aid this effort, EHDEN held a Rapid Collaboration Call inviting institutions across Europe with COVID-19 data to participate [12]. This presented a challenge: how can we be sure the data are of high quality for research while quickly integrating them into the network?

EHDEN proposes to solve this problem of data integration by choosing to harmonize the network on the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [13]. European real-world clinical data is generated from diverse medical record systems, stored in different ways, captured in different languages, and controlled by differing policy restrictions. Conversion to the OMOP CDM allows for interoperability across data sources as it standardizes real world data to both a common structure (data model) and terminology (vocabulary). This then leaves the problem of ensuring these data are of high-quality to support evidence generation.

While there is little doubt in the research utility of observational health data, concerns are often voiced about the quality of the data since it is not primarily collected for research purposes. A study published in 2018 by Pacurariu et al. detailed a review of 34 observational health databases representing Northern, Central, and Western Europe [14]. The authors found that only half had published validation studies detailing the extent to which patients' records aligned with national registries or statistics reports [14]. Among databases studied, the approaches taken to assess data quality were observed to be very

different. Clinical Practice Research Datalink (CPRD), for example, provides indicators of patient and practice acceptability to researchers [15]. In contrast, the Information System for the Development of Research in Primary Care (SIDIAP) from Catalonia, Spain created a scoring system to assess the completeness of the data provided by primary care practitioners [16].

To address this lack of standard data quality procedures, EHDEN, in conjunction with the Observational Health Data Sciences & Informatics (OHDSI) initiative [17,18], developed a novel process for assessing the quality of observational health data, known as the Data Quality Dashboard (DQD) (https://ohdsi.github.io/DataQualityDashboard/, accessed on 20 September 2021) [13,19,20]. While the DQD has proven effective when applied to a singular database, it has not yet been applied to a network of databases such as those within EHDEN. For this work, the COVID-19 Rapid Collaboration Call gave us the unique opportunity to study how the use of a standard data quality procedure improves the quality of data across a federated network.

With this work herein we present the process by which Data Partners (DPs) chosen for the COVID-19 Rapid Collaboration Call mapped their data to the OMOP CDM. Throughout the process, at least two DQD results were collected and compared for each DP to understand how the tool improves the quality of observational health databases at scale. This is the first study to apply the DQD to multiple, disparate databases representing different countries and data types, and is an important step forward towards transparent data provenance.

2. MATERIALS AND METHODS

2.1. Data Conversion Process and DQD Collection

Twenty-five DPs were awarded a COVID-19 Rapid Collaboration Call grant and these DPs covered 11 different countries and represented over 1 million COVID-19 patients [12]. The data sources themselves ranged from small COVID-19 registries to large medical records systems that covered lives for nearly an entire country. Once a DP was notified that they would receive a grant, they were assigned an EHDEN COVID-19 Taskforce. This taskforce consisted of a small team of technical experts with the collective skills and experience to walk DPs through the Data Conversion to Analysis Pipeline shown in Figure 1.

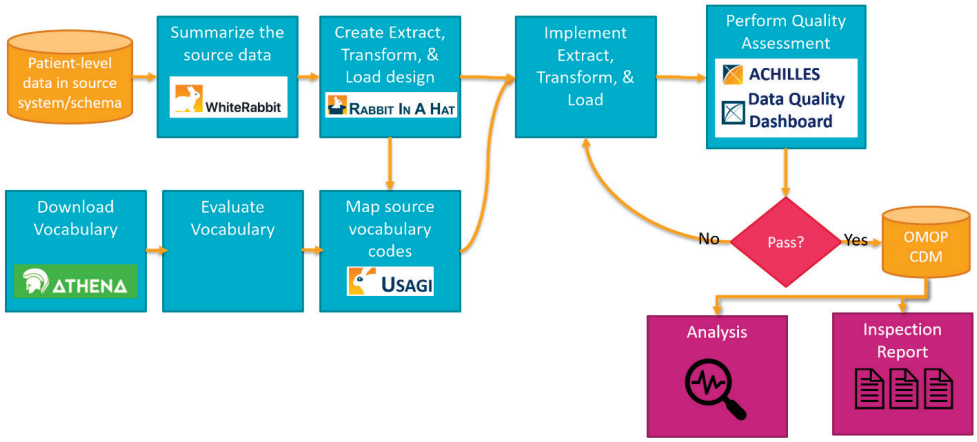


Figure 1: The Data Conversion to Analysis Pipeline to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) used by the European Health Data and Evidence Network (EHDEN).

This Data Conversion to Analysis Pipeline begins with WhiteRabbit and Rabbit-In-A-Hat [21], two tools maintained by EHDEN and used by the global OHDSI community. These tools facilitate the structural standardization of data by first revealing what tables, fields, and values are in a source data set (WhiteRabbit) and then allowing users to interactively map those source tables and fields to the CDM by way of a graphical user interface (Rabbit-In-A-Hat). Once the structural design specification is complete, standardization of the medical terminologies present in a source dataset can be done in two ways. First, one can leverage existing translations of common source terminologies (e.g., ICD10, Read, etc.) to standard concepts using the OMOP Vocabularies [22,23]. If the source data contains codes not found in the OMOP Vocabularies, a tool called Usagi [24] is used to suggest potential standard concept mappings, followed by manual review and correction. After designing the structural and semantic standardization, Rabbit-In-A-Hat produces a document containing the complete instructions to convert the source data to an OMOP CDM instance. DPs then implement these specifications in an ETL (Extract, Transform, & Load) program utilizing the hardware and technologies they have access to within their organization. Typically, this takes the form of SQL written in the database environment they already have.

Once data has been standardized, the next part of the Data Conversion to Analysis Pipeline is quality assessment, the cornerstone of which within EHDEN is the DQD [19,20]. DQD is an open-source R package supporting OMOP CDM versions 5.2.2 and 5.3.1 that uses a systematic approach to run data quality checks against an OMOP CDM instance. It works with multiple database management systems, including PostgreSQL, Microsoft SQL Server, and Amazon Redshift. This tool applies, to a given database, over 3300 data quality checks (Table A1) organized into categories first described by Kahn et al. [25].

These are conformance (requiring data to adhere to specified standards and formats), completeness (ensuring that data values are present), and plausibility (ensuring that the data values are believable). Using the MEASUREMENT table and the MEASUREMENT_CONCEPT_ID column as a motivating example, a conformance check ensures the column exists, a completeness check ensures it is non-zero above a given threshold, and a plausibility check for a given measurement (e.g., Calcium; total) and given unit (e.g., milligram per deciliter) ensures the value is biologically plausible.

The quality checks in the DQD tool can be described at a high-level by their check type. There are 20 types in total, the full list of which can be found in Appendix A. Each type represents a different data quality idea which is then automatically resolved and applied to the relevant tables and fields with data available as the DQD is run. These check types range in complexity from checking relational constraints to evaluating plausible values for measurement-unit pairs. For example, cdmField, a conformance quality check, is a check type looking to see if all fields are present in the CDM as expected and for each field in this check type there is a result of either pass or fail based on if the field is correctly implemented.

Once the checks are applied and run, they each return the number of rows that failed the logical test and the total number of rows eligible to be inspected for the given quality metric. Dividing the number of failing rows by the total number of rows gives the proportion of failing rows for each data quality check. This proportion is then compared to a pre-specified threshold. If the proportion of failing rows exceeds the threshold, then the check is considered to have failed.

Once a DP completed the first iteration of their standardized database, they ran the DQD in their environment and produced the data quality check results. These results were then reviewed with the help of the EHDEN COVID-19 Taskforce using the DQD results viewer. An example of what the results viewer looks like can be found at https://data.ohdsi.org/DataQualityDashboardMDCD/ (accessed on 20 September 2021). Together, they determined whether the issues identified by the tool were most likely related to the source data or the ETL program. In the case of a source data problem the DPs would fix the issue, if possible, usually by instituting a rule in the ETL (e.g., dropping patients without a recorded year of birth). Occasionally issues were fixed at the source itself, such as the DP for whom everyone with a death date also had encounter records after death. This turned out to be a problem with their electronic health record system, which they went back and corrected.

In the case of a problem in the ETL program the EHDEN taskforce would work with the DP to pinpoint and repair the bugs in the code. If an issue could not be immediately fixed an explanation was written in the DQD tool clearly describing the reason for the check

failure. After the issues in the source data and the ETL code were addressed, the DQD was run again. This process of running DQD and updating the ETL, source data, or DQD configuration continued until the taskforce felt confident that the database was of sufficient quality, as seen in Figure 2.

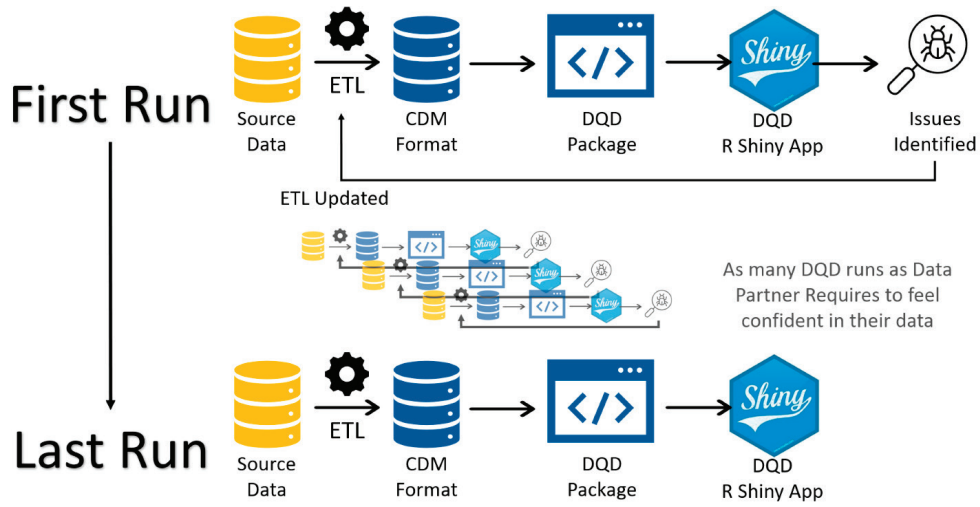


Figure 2: The Extract, Transform, & Load (ETL) to Common Data Model (CDM) to Data Quality Dashboard (DQD) Feedback Loop used in the European Health Data & Evidence Network (EHDEN) COVID-19 Rapid Collaboration Call.

2.2. Data Quality Dashboard Comparisons

At the end of the entire data conversion process, when a DP achieved an OMOP CDM-formatted database they and the taskforce were confident in, both the first and last DQD reports were collected. The data quality checks comprising these reports were then organized into four outcome categories: a *pass* indicated that the percentage of failing rows for a given quality check was below the threshold specified in the DQD tool, a *fail* indicated that the percentage of failing rows for a given quality check was above the threshold specified, an *error* indicated that there was an error in the SQL and no result for the given quality check was returned, and *no data* indicated that there was no data in the database that satisfied the criteria for a given quality check.

# of passed checks(# of passed checks+# of failed checks+# of check errors)×100# of passed checks(# of passed checks+# of failed checks+# of check errors)×100

After organizing the data quality checks by outcome, the results from the first DQD report were compared with the results from the last DQD report. Initially, a simple check pass percentage (e.g., 95% of the DQD checks passed) was obtained by DP and time (first or

last) using Equation (1) to determine if the percent of data quality checks passed in the last run of the DQD was higher than the percent of data quality checks passed in the first run of the DQD. It was then observed that almost all DPs ran a different number of quality checks at first versus at last. Upon inspection it was revealed that this phenomenon was due to a few factors: the iterative ETL process, the automatic nature of the tool, and the fact that DPs removed checks with no data to support them from the calculation. For example, many DPs entered the Data Conversion to Analysis Pipeline beginning with the information they believed was easiest to standardize, typically patient demographics and patient diagnoses. Once they achieved the first standardized database containing this information, many ran the DQD with only a subset of the total checks turned on. These tended to be checks that evaluated conformance to the model and completeness of standard vocabulary mapping. As they iterated through the process, the DPs continued to turn on more checks, usually those evaluating plausibility. To account for this, and to enable more appropriate comparisons, we conducted a second analysis limiting to only those data quality checks that were run both during the first DQD and the last DQD.

As mentioned above, the DQD determines a pass or a fail by comparing the percentage of rows that fail a check to a pre-specified threshold. While this gives an immediate idea of how well a quality check performed in a database, it only provides part of the story. The binary nature of assigning a pass or fail label overlooks the notion of improvement. Since we were interested in how well the DPs improved the quality of their database, we quantified, by DP, the number of checks that decreased the percent failing rows from first to last DQD run in addition to the passes and fails.

To gain insight into which types of checks failed most often in the beginning compared to the end of the data conversion process we also stratified the percent of passing checks by check type and time (first or last).

3. RESULTS

At the time of this writing, 15 of the 25 DPs completed their DQD review, 8 had not completed, and 2 were not considered, as they did not run DQD for various reasons. Table 1 lists, in alphabetical order, the 15 DPs that contributed data for this work. These 15 DPs represented 10 countries, primarily were electronic medical record (EHR) data, and ranged in size from 766 subjects to 2.8 M subjects. The types of subjects captured within each DP were diverse in type (e.g., general practitioner data versus hospital data), depth (e.g., data only about COVID-19 care to data on all clinical care) and complexity (e.g., 1 table to 100s of tables). These DPs also represented diverse technology platforms that include several types of database management systems (the most popular being PostgreSQL, Microsoft SQL Server, and Amazon Redshift), differing infrastructure set up (e.g., on premise versus

in the cloud), and differing number of processors (1 to 56) and memory sizes (8 GB to 1.5 TB) at their disposal. Additional information about each DP can be found at https://www.ehden.eu/datapartners/ (accessed on 1 October 2021). For the rest of this paper, these DPs will be anonymized.

Table 1: Data Partners (DPs) that contributed data (alphabetical order).

Data Partner	Data Type	Country	Total No of Subjects
Amsterdam University Medical Centers—COVID-19 ICU Decision Support (Amsterdam UMC) *	EHR	The Netherlands	1.8 K
Clinical Practice Research Datalink (CPRD) AURUM	EHR	United Kingdom	409 K **
Danish Colorectal Cancer Group (DCCG) with Center for Surgical Science	Patient Registry	Denmark	76.8 K
Fondazione IRCCS Istituto Neurologico Carlo Besta COVID-19 Database (FINCB) *	Patient Registry	Italy	766
Health Data Warehouse of Assistance Publique—Hopitaux de Marseille (AP-HM)	EHR	France	2.5 M
Health Informatics Centre (HIC)	EHR	United Kingdom	1.2 M
Heliant—University Clinical Center of Serbia (CCSERBIA)	EHR	Serbia	860 K
Information System for Research in Primary Care—Hospitalization Linked Data (SIDIAP)	EHR	Spain	7.8 M
Istanbul University—Faculty of Medicine (IU) *	EHR	Turkey	59.6 K
LynxCare	EHR	Belgium	10 K
Medaman Hospital Data (MHD)	EHR	Belgium	104 K
Parc de Salut Mar Barcelona and Hospital del Mar Medical Research Institute (IMIM)	EHR	Spain	956 K
Servicio Cántabro de Salud and IDIVAL (IDIVAL)	EHR	Spain	580 K
UK Biobank via University College London Institute of Health Informatics	Biobank	United Kingdom	500 K
Unidade Local de Saúde de Matosinhos (ULSM) COVID-19 *	EHR	Portugal	9.7 K

EHR = electronic health record; * = Common Data Model only contain COVID-19 subjects, ** = small subset of data for conversion purposes.

For each of the 15 DPs, the minimum number of DQD runs was 2 times, maximum 33, with a median of 4 times. The total execution time on the first run of the DQD ranged from 0.02 h to 10.5 h with a median of 0.7 h. The total execution time on the last run of the DQD ranged from 0.02 h to 25.0 h with a median of 0.7 h. These ranges were heavily dependent on the size of the database, the size of the server the DP used, and the amount of compute available. The duration of time over which the first and last DQD were run ranged from 10 days to 190 days, with a median of 90 days. The number of DQD runs and time to completion were impacted by many things such as the number of issues

identified by DQD that needed to be fixed, desire to fix all issues identified by DQD, the amount of time a specific DP could afford to spend working on their ETL conversion, and experience with ETL conversions and the ETL technology required to do so (thus higher number of subjects did not necessarily equate to more DQD issues as there are many factors at play).

After limiting the comparison to the checks that were run both at first and at last, all DPs had a higher percent passing in the last run of the DQD compared to the first (Table 2), with a median of 883 total checks evaluated. We also quantified the number of checks that saw a reduction in the percent of failing rows even though they still failed the DQD check. Most DPs only had 1 or 2 checks that were labeled as failures though they showed improvement, except for DP8 and DP9, which had 12 and 8 respectively.

When expanded to all checks run (not just those in common between the first and last) and stratified by check type, we found that more total checks were evaluated in the last run for all check types except cdmField and measurePersonCompleteness (Table 3). With this increase in total checks, we also saw an increase in the percent of passing checks for all check types from first to last.

We found that the data quality checks that failed most often in the first run of the DQD were related to checks ensuring that the database conforms to the specifications of the OMOP CDM. The worst offenders in the first run were the checks in the fkclass and fkdomain check types, which are conformance checks reviewing if the classes and domains of mapped terminology are appropriate. Initially only 66.7% passed for fkclass and 70.8% passing for fkdomain. While overall the conformance checks failed most often in the first run of the DQD, they also showed the most improvement from first to last, achieving between 100% and 96.5% passing at last run.

The percent of passing completeness and plausibility checks also increased from first to last run of the DQD, but not as drastically as the conformance checks. The most improved completeness checks belonged to the measurePersonCompleteness check type, going from 81.5% passing at first to 96% passing at last. Similarly, the most improved plausibility checks belonged to the plausibleDuringLife check type, going from 70.5% passing at first to 92% passing at last.

Table 2: The proportion of data quality checks that either passed or reduced failing rows between the first and last run of Data Quality Dashboard (DQD), by Data Partner (DP), for checks in common between the two.

Data Partner	First DQD Run			Last DQD Run			
	Passing Checks	Number of Checks	% Passing	Passing Checks	Checks with Failing Rows	Number of Checks	% Passing or Decrease Failing Rows
1	981	988	99.3%	984	0	988	99.6%
2	677	696	97.3%	678	1	696	97.6%
3	712	723	98.5%	721	0	723	99.7%
4	738	963	76.6%	911	2	963	94.8%
5	799	925	86.4%	916	0	925	99.0%
6	1015	1055	96.2%	1046	2	1055	99.3%
7	872	877	99.4%	875	0	877	99.8%
8	744	773	96.2%	751	12	773	98.7%
9	922	959	96.1%	942	8	959	99.1%
10	697	803	86.8%	802	0	803	99.9%
11	778	883	88.1%	826	2	883	93.8%
12	668	696	96.0%	687	1	696	98.9%
13	858	952	90.1%	937	4	952	98.8%
14	949	1044	90.9%	1028	2	1044	98.7%
15	822	871	94.4%	861	0	871	98.9%

Table 3: The proportion of passing data quality checks between the first and last run of Data Quality Dashboard (DQD), for all checks run, by high-level check type and Summarized across all Data Partners (DPs)

Kahn Category	Check Type	First DQD Run			Last DQD Run		
		Passing Checks	Number of Checks	% Passing	Passing Checks	Number of Checks	% Passing
Conformance	FKCLASS	8	12	66.7%	14	14	100.0%
Conformance	CDMDATATYPE	1143	1147	99.7%	1268	1268	100.0%
Conformance	ISREQUIRED	740	768	96.4%	853	854	99.9%
Conformance	CDMFIELD	4215	4341	97.1%	4011	4017	99.9%
Conformance	ISPRIMARYKEY	159	162	98.1%	183	185	98.9%
Conformance	ISSTANDARDVALIDCONCEPT	344	407	84.5%	415	423	98.1%
Conformance	ISFOREIGNKEY	862	1116	77.2%	1097	1119	98.0%
Conformance	FKDOMAIN	235	332	70.8%	328	340	96.5%
Completeness	MEASUREVALUECOMPLETENESS	2244	2313	97.0%	2499	2501	99.9%
Completeness	SOURCECONCEPTRECORD COMPLETENESS	128	141	90.8%	142	146	97.3%
Completeness	SOURCEVALUECOMPLETENESS	237	254	93.3%	271	281	96.4%
Completeness	MEASUREPERSONCOMPLETENESS	176	216	81.5%	193	201	96.0%
Completeness	STANDARDCONCEPTRECORD COMPLETENESS	327	381	85.8%	377	416	90.6%
Plausibility	PLAUSIBLEGENDER	1017	1228	82.8%	1410	1481	95.2%
Plausibility	PLAUSIBLEVALUELOW	383	435	88.0%	584	618	94.5%
Plausibility	PLAUSIBLETEMPORALAFTER	265	334	79.3%	355	377	94.2%
Plausibility	PLAUSIBLEDURINGLIFE	146	207	70.5%	242	263	92.0%
Plausibility	PLAUSIBLEVALUEHIGH	372	406	91.6%	514	594	86.5%

4. DISCUSSION

Throughout the COVID-19 Rapid Collaboration Call, selected DPs mapped their data to the OMOP CDM and applied the DQD at least twice along the way. These participants were from different countries with different types of data, all working toward the goal of developing a research-ready standardized database. Multiple DQD results per DP allowed for comparisons to understand how standard data quality procedures improve the quality of a federated network. When limited to checks that were in common between the first and last run of the DQD, we saw an increase in the percent of passing data quality checks for all DPs. In other words, the DQD helped all DPs improve their data quality.

Moreover, when we expanded our scope to look at all checks that were run, stratifying on check type, we found that a higher number of checks were run at last then at first. The only two check types that did not follow this pattern were cdmField and measurePersonCompleteness because some DPs turned off checks for tables that did not have data. Equation (1) removes almost all checks without data to support them from the percent passing calculation; however, based on how the cdmField and measurePersonCompleteness check types function, they would still report a result even if there was no data in the table. Turning off the checks in the tool for these tables with no data helped the DPs to remove failures that were uninformative for their data.

As the DPs continued to iterate through the process and run the DQD, the increase in the total number of checks run tells us that the tool verified the past issues and then continued to check new items. These new items were a result of either the DPs turning on checks they intentionally turned off in the first run, they were a result of continued vocabulary mapping, or additional data being added to the OMOP CDM. The more proprietary source codes were mapped to standard concepts, the more checks the DQD could leverage against those records. Specifically, these are checks evaluating measurement plausibility as they are dependent on individual concepts representing individual source codes. Ultimately, the DQD became more and more comprehensive the better the DPs got at mapping their data to the OMOP CDM.

We also found that the conformance checks improved the most from first to last as compared with the completeness and plausibility checks. The DQD seemed the most effective at iterating on and helping to improve those checks that assess the conformance to the model. While not perfect at the last run, all conformance checks achieved a percent passing of 96.5% or higher, with two achieving 100%. The handful of primary key and foreign key checks that did not pass during the last DQD run were found to be related to the VISIT_OCCURRENCE_IDs for two DPs. These are required to be unique in the VISIT_OCCURRENCE table. One DP was perfect on the first run but showed a handful of duplicates in the last run. Most likely this was due to the increase in records from first to last; the ETL

performed well on a small subset but introduced duplicates when expanded to the full database. The other DP showed great improvement on this issue between first and last, going from 1.3 M to 4 K offending records. Since it was not 100% perfect in the last run it was still labelled as a failure though the DP clearly worked to address the problem.

Viewed from the network level this focus on adherence to model specifications is extremely important. A federated network functions with each data holder maintaining their OMOP CDM compliant database behind a firewall. They participate in studies by running standardized code against their OMOP CDM instance and reporting the results. Therefore, if the database does not conform to the model, they would not be able to run the study code, let alone report results.

This was especially evident as DPs easily ran a network-based research study after the quality assessment was completed. When the first handful of DPs achieved a standardized database both they and the taskforce were confident in, they were invited to participate in an ongoing research study on adverse events of special interest (AESI) for the COVID-19 vaccines [26,27]. Thanks to the data quality review all DPs were able to successfully run the package without any major conformance issues hindering their involvement.

While the DQD functioned well to help DPs address issues of conformance, we saw less improvement in the completeness and plausibility checks. There was still an increase in the percent of passing checks from first to last run, but it was less dramatic than that of the conformance checks. This was partly due to the amount of work needed to improve each one. Completeness, in terms of the OMOP CDM, can mean data completeness but also completeness of standard vocabulary mapping. If a DP has a large source vocabulary that needs to be mapped to standard terminologies, this work can take time to address, and it is only once hundreds or thousands of codes are mapped to standard concepts that the data quality check moves from a fail to a pass.

In addition to the amount of work required, we often found it difficult to assess the completeness and plausibility checks when considering the entire database. The number of records that needed to be addressed was daunting and there was no way to know which ones would potentially impact an analysis. Therefore, our recommendation is to institute not only database-level data quality checks but study-specific checks. It is important to know the database is conformant enough to run a study, but once divided into cohorts, the field of focus for data quality issues is narrowed and may reveal items that were masked at the higher level. For example, when running the AESI COVID-19 vaccine study, one DP found that most of the patients in the defined study cohort did not have visit information. However, another discovered that many of the lab values of interest for the analysis did not have units of measure mapped to standard concepts. Both of these items were identified during execution of the study, addressed, and then the study was re-run.

5. CONCLUSIONS

This is the first study to date that demonstrates the DQD, as an analytic tool, can be applied to a diverse collection of databases representing different types of data across many countries. The DQD improved the quality of all participating databases, rendering them ready for analysis. We recommend this practice to all data holders either working to convert their data to the OMOP CDM or who already have an OMOP CDM instance. While study-specific quality checks should still be performed, the DQD ensures conformance to the model specifications and that a database meets a baseline level of completeness and plausibility for use in research.

AUTHOR CONTRIBUTIONS

Conceptualization, C.B., E.A.V., M.J.S. and P.R.R.; Data curation, C.B. and E.A.V.; Formal analysis, C.B. and E.A.V.; Visualization, E.A.V. and F.D.; Writing—original draft, C.B. and E.A.V.; Writing—review & editing, F.D., N.H., M.J.S., M.M. and P.R.R. All authors have read and agreed to the published version of the manuscript.

DECLARATIONS

Funding

This research was funded by the European Union's Horizon 2020 research and innovation programme and EFPIA, grant number 806968. The APC was funded by Janssen Research & Development, LLC.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Acknowledgments

The authors want to thank the data partners who participated in this research and the Innovative Medicines Initiative.

Conflicts of Interest

C.B., E.A.V., F.D., N.H. and M.J.S. are employees of Janssen Research & Development, LLC and Johnson & Johnson shareholders. M.M. and P.R.R. has no conflicts to declare.

APPENDIX A.

Data Quality Check Types (Ideas) by Context and Category

The information within this table can additionally be found here: https://ohdsi.github.io/DataQualityDashboard/articles/CheckTypeDescriptions.html.

Table A1: Kahn category**.**

Check Name	Check Description	Kahn Category [1]
cdmField	A yes or no value indicating if all fields are present in the @cdmTableName table as expected based on the specification.	Conformance
isRequired	The number and percent of records with a NULL value in the @cdmFieldName of the @cdmTableName that is considered not nullable.	Conformance
cdmDatatype	A yes or no value indicating if the @cdmFieldName in the @cdmTableName is the expected data type based on the specification.	Conformance
isPrimaryKey	The number and percent of records that have a duplicate value in the @cdmFieldName field of the @cdmTableName.	Conformance
isForeignKey	The number and percent of records that have a value in the @cdmFieldName field in the @cdmTableName table that does not exist in the @fkTableName table.	Conformance
fkDomain	The number and percent of records that have a value in the @cdmFieldName field in the @cdmTableName table that do not conform to the @fkDomain domain.	Conformance
fkClass	The number and percent of records that have a value in the @cdmFieldName field in the @cdmTableName table that do not conform to the @fkClass class.	Conformance
isStandardValidConcept	The number and percent of records that do not have a standard, valid concept in the @cdmFieldName field in the @cdmTableName table.	Conformance
measureValueCompleteness	The number and percent of records with a NULL value in the @cdmFieldName of the @cdmTableName.	Completeness
measurePersonCompleteness	The number and percent of persons in the CDM that do not have at least one record in the @cdmTableName table	Completeness
standardConceptRecordCompleteness	The number and percent of records with a value of 0 in the standard concept field @cdmFieldName in the @cdmTableName table.	Completeness
sourceConceptRecordCompleteness	The number and percent of records with a value of 0 in the source concept field @cdmFieldName in the @cdmTableName table.	Completeness
sourceValueCompleteness	The number and percent of distinct source values in the @cdmFieldName field of the @cdmTableName table mapped to 0.	Completeness
plausibleValueLow	The number and percent of records with a value in the @cdmFieldName field of the @cdmTableName table less than @plausibleValueLow.	Plausibility
plausibleValueHigh	The number and percent of records with a value in the @cdmFieldName field of the @cdmTableName table greater than @plausibleValueHigh.	Plausibility

Table A1: Kahn category**.** (*continued*)

Check Name	Check Description	Kahn Category [1]
plausible TemporalAfter	The number and percent of records with a value in the @cdmFieldName field of the @cdmTableName that occurs prior to the date in the @plausibleTemporalAfterFieldName field of the @plausibleTemporalAfterTableName table.	Plausibility
plausible DuringLife	If yes, the number and percent of records with a date value in the @cdmFieldName field of the @cdmTableName table that occurs after death.	Plausibility
plausibleValueLow	For the combination of CONCEPT_ID @conceptId (@conceptName) and UNIT_CONCEPT_ID @unitConceptId (@unitConceptName), the number and percent of records that have a value less than @plausibleValueLow.	Plausibility
plausibleValueHigh	For the combination of CONCEPT_ID @conceptId (@conceptName) and UNIT_CONCEPT_ID @unitConceptId (@unitConceptName), the number and percent of records that have a value higher than @plausibleValueHigh.	Plausibility
plausibleGender	For a CONCEPT_ID @conceptId (@conceptName), the number and percent of records associated with patients with an implausible gender (correct gender = @plausibleGender).	Plausibility

[1] Kahn, M.G.; Callahan, T.J.; Barnard, J.; Bauck, A.E.; Brown, J.; Davidson, B.N.; Estiri, H.; Goerg, C.; Holve, E.; Johnson, S.G.; et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)* **2016**, *4*, 1244, doi:10.13063/2327-9214.1244.

REFERENCES

1. Mitchell, J.B.; Bubolz, T.; Paul, J.E.; Pashos, C.L.; Escarce, J.J.; Muhlbaier, L.H.; Wiesman, J.M.; Young, W.W.; Epstein, R.S.; Javitt, J.C. Using Medicare claims for outcomes research. Med. Care 1994, 32, 38–51.
2. Lewis, N.J.W.; Patwell, J.T.; Briesacher, B.A. The Role of Insurance Claims Databases in Drug Therapy Outcomes Research. PharmacoEconomics 1993, 4, 323–330.
3. Adler-Milstein, J.; DesRoches, C.M.; Kralovec, P.; Foster, G.; Worzala, C.; Charles, D.; Searcy, T.; Jha, A.K. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff. 2015, 34, 2174–2180.
4. Scherer, D.G.; Annett, R.D.; Brody, J.L. Ethical Issues in Adolescent and Parent Informed Consent for Pediatric Asthma Research Participation. J. Asthma 2007, 44, 489–496.
5. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. J. Big Data 2019, 6, 54.
6. Gershon, A.S.; Lindenauer, P.K.; Wilson, K.C.; Rose, L.; Walkey, A.J.; Sadatsafavi, M.; Anstrom, K.J.; Au, D.H.; Bender, B.G.; Brookhart, M.A.; et al. Informing Healthcare Decisions with Observational Research Assessing Causal Effect. An Official American Thoracic Society Research Statement. Am. J. Respir. Crit. Care Med. 2021, 203, 14–23.
7. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. Available online: https://www.fda.gov/media/120060/download (accessed on 12 October 2021).
8. Heads of Medicines Agency. European Medicines Agency HMA-EMA Joint Big Data Taskforce Summary Report. Available online: https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf (accessed on 12 October 2021).
9. IMI2–Call 12. Available online: http://www.imi.europa.eu/apply-funding/closed-calls/imi2-call-12 (accessed on 15 October 2021).
10. European Health Data Evidence Network. Available online: https://www.ehden.eu/ (accessed on 12 March 2021).
11. Hrabovszki, G. COVID-19: Latest Updates. Available online: https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/covid-19-latest-updates (accessed on 15 October 2021).
12. 04/2020—COVID19 Rapid Collaboration Call. Available online: https://www.ehden.eu/open-calls/04-2020-covid19-data-partner-call/ (accessed on 20 September 2021).
13. OMOP Common Data Model. Available online: http://ohdsi.github.io/CommonDataModel/ (accessed on 14 January 2021).
14. Pacurariu, A.; Plueschke, K.; Mcgettigan, P.; Morales, D.R.; Slattery, J.; Vogl, D.; Goedecke, T.; Kurz, X.; Cave, A. Electronic healthcare databases in Europe: Descriptive analysis of characteristics and potential for use in medicines regulation. BMJ Open 2018, 8, e023090.
15. Herrett, E.; Gallagher, A.M.; Bhaskaran, K.; Forbes, H.; Mathur, R.; van Staa, T.; Smeeth, L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int. J. Epidemiol. 2015, 44, 827–836.
16. Del Mar García-Gil, M.; Hermosilla, E.; Prieto-Alhambra, D.; Fina, F.; Rosell, M.; Ramos, R.; Rodriguez, J.; Williams, T.; van Staa, T.; Bolibar, B. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform. Prim. Care 2011, 19, 135–145.
17. Observational Health Data Sciences & Informatics (OHDSI). Available online: www.ohdsi.org (accessed on 20 September 2021).

18. Hripcsak, G.; Duke, J.D.; Shah, N.H.; Reich, C.G.; Huser, V.; Schuemie, M.J.; Suchard, M.A.; Park, R.W.; Wong, I.C.K.; Rijnbeek, P.R.; et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Re-searchers. Study Health Technol. Inform. 2015, 216, 574–578. [Google Scholar]
19. Blacketer, C.; Defalco, F.J.; Ryan, P.B.; Rijnbeek, P.R. Increasing trust in real-world evidence through evaluation of observational data quality. J. Am. Med. Inform. Assoc. 2021, 28, 2251–2257.
20. DataQualityDashboard (DQD). Available online: https://ohdsi.github.io/DataQualityDashboard/ (accessed on 10 October 2021).
21. WhiteRabbit. Available online: https://github.com/OHDSI/WhiteRabbit (accessed on 20 September 2021).
22. Vocabulary. Available online: https://github.com/OHDSI/Vocabulary-v5.0 (accessed on 12 October 2021).
23. The Book of OHDSI. Available online: https://ohdsi.github.io/TheBookOfOhdsi/ (accessed on 22 September 2021).
24. Usagi. Available online: https://github.com/OHDSI/Usagi (accessed on 22 September 2021).
25. Kahn, M.G.; Callahan, T.J.; Barnard, J.; Bauck, A.E.; Brown, J.; Davidson, B.N.; Estiri, H.; Goerg, C.; Holve, E.; Johnson, S.G.; et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Gener. Évid. Methods Improv. Patient Outcomes) 2016, 4, 18–1244.
26. Releases OHDSI/CommonDataModel. Available online: https://github.com/OHDSI/CommonDataModel/releases (accessed on 10 October 2021).
27. Calculating the Background Rates of Adverse Events of Special Interest (AESI) for the COVID Vaccines. Available online: https://github.com/ohdsi-studies/Covid19VaccineAesiIncidenceRate (accessed on 18 October 2021).

Fitness-for-Purpose

5

Building the Largest Federated Health Data Network in Europe: Advancing Collaboration and Real-World Evidence Generation

Clair Blacketer[1,2,5]
Martijn J. Schuemie[1,5,6]
Maxim Moinat[1,2]
Erica A Voss[1,5]
Montse Camprubi[1,3]
Peter R. Rijnbeek[1,2]*
Patrick B. Ryan[1,4,5]*

[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]SYNAPSE Research Management Partners, Madrid, ES
[4]Department of Biomedical Informatics, Columbia University, New York, NY, USA
[5]Johnson & Johnson, Raritan, NJ, USA
[6]Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA, USA

Blacketer C, Schuemie MJ, Moinat M, Voss EA, Camprubi M, Rijnbeek PR, Ryan PB. Advancing Real-World Evidence Through a Federated Health Data Network (EHDEN): Descriptive Study. J Med Internet Res. 2025 Aug 7;27:e74119. doi: 10.2196/74119

ABSTRACT

Background

Real-world data (RWD) are increasingly used in health research and regulatory decision-making to assess the effectiveness, safety, and value of interventions in routine care. However, the heterogeneity of European healthcare systems, data capture methods, coding standards, and governance structures poses challenges for generating robust and reproducible real-world evidence (RWE). The European Health Data & Evidence Network (EHDEN) was established to address these challenges by building a large-scale, federated data infrastructure that harmonizes RWD across Europe.

Objective

This study aims to describe the composition and characteristics of the databases harmonized within EHDEN as of September 2024. We seek to provide transparency regarding the types of RWD available and their potential to support collaborative research and regulatory use.

Methods

EHDEN recruited data partners (DPs) through structured open calls. Selected DPs received funding and technical support to harmonize their data to the OMOP Common Data Model (CDM), with assistance from certified small-to-medium enterprises (SMEs) trained through the EHDEN Academy. Each data source underwent an extract-transform-load (ETL) process and data quality assessment using the Data Quality Dashboard. Metadata—including country, care setting, capture method, and population criteria—were compiled in the publicly accessible EHDEN Portal.

Results

As of September 1, 2024, the EHDEN Portal includes 210 harmonized data sources from 30 countries. The highest representation comes from Italy (13.0%), Great Britain (12.5%), and Spain (11.5%). The mean number of persons per data source is 2,147,161, with a median of 457,664 individuals. Regarding care setting, 46.7% (n=98) of data sources reflect data exclusively from secondary care, 42.4% (n=89) from mixed care settings (both primary and secondary), and 11.0% (n=23) from primary care only.

In terms of population inclusion criteria, 55.7% (n=117) of data sources include individuals based on healthcare encounters, 32.9% (n=69) through disease-specific data collection, and 11.4% (n=24) via population-based sources. Data capture methods also vary, with electronic health records (EHRs) being the most common. A total of 74.7% (n=157) of data sources utilize EHRs, and more than half of those (n=85) rely on EHRs as their sole method of data collection. Laboratory data are used in 29.5% (n=62) of data sources,

though only one relies exclusively on lab data. Most laboratory-based data sources combine this method with other forms of data capture.

Conclusion

EHDEN is the largest federated health data network in Europe, enabling standardized, GDPR-compliant analysis of RWD across diverse care settings and populations. This descriptive summary of the network's data sources enhances transparency and supports broader efforts to scale federated research. These findings demonstrate EHDEN's potential to enable collaborative studies and generate trusted evidence for public health and regulatory purposes.

INTRODUCTION

Real-world data (RWD) has become a cornerstone in healthcare research, especially in regulatory science, due to its ability to capture insights from diverse patient populations and clinical settings. Unlike data generated through traditional randomized controlled trials (RCTs), which often have stringent inclusion criteria, RWD reflects the everyday healthcare experiences of a broader patient base[1–5]. This breadth offers a richer context for understanding drug safety and effectiveness, guiding post-authorization safety monitoring, informing risk-benefit evaluations, and supporting regulatory decisions[6]. Regulators, industry, and academics alike rely on real-world evidence (RWE) derived from RWD to answer critical questions about healthcare interventions in clinical care settings that are more representative of routine practice[7–9].

Europe's healthcare landscape presents both challenges and opportunities for generating RWD[10]. Its diversity spans many different health systems, terminology systems, and data collection practices, with variability in healthcare delivery and data availability across countries. This heterogeneity complicates large-scale representative research but also offers a unique opportunity to study diverse populations[8,11–13]. However, capturing this potential requires overcoming technical, operational, and methodological barriers to ensure data harmonization and quality. A federated network is particularly well suited to Europe's fragmented healthcare landscape, where legal, linguistic, and governance diversity necessitate a model that supports local control while enabling cross-border collaboration.

Federated data networks, like the European Health Data & Evidence Network (EHDEN), are well-suited for Europe's decentralized data landscape[14–19]. In the context of EHDEN, a federated network refers to a collaboration of independently governed data sources that retain full control of their data locally, preserving the autonomy and governance policies of individual data holders. This approach allows for multi-database studies across diverse populations without requiring centralized data access or query execution, ensuring that personal health information does not leave its original source. By design, this method complies with the General Data Protection Regulation (GDPR), as it avoids centralizing or transferring personal data and supports the principles of data minimization, purpose limitation, and local control. It is important to note that privacy-preserving practices are also in place for studies conducted using the network. Data Partners (DPs) only share aggregate results – typically high-level outputs such as hazard ratios – after applying a minimum cell size threshold (k-anonymity, commonly set to 5) to suppress potentially re-identifiable results.

EHDEN was established as an Innovative Medicines Initiative (IMI), now Innovative Health Initiative (IHI) public-private partnership in November 2018 to overcome the challenges

and transform how health data is utilized in Europe[20,21]. The project built a federated data network that standardizes health data across participating sources, making data analysis more feasible and consistent. By harmonizing data and implementing quality assurance protocols, EHDEN enhances the usability and comparability of real-world data across Europe. This paper provides an overview of the EHDEN network, examining its data harmonization efforts, quality control processes, and the range of data sources included in the network. Through this discussion, we aim to highlight the scope of RWD available across Europe and its potential for advancing healthcare research and regulatory decision-making.

METHODS

Common Data Model

As the foundation for its network, EHDEN adopted the OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM)[22–24]. The OMOP CDM is widely recognized for its "structure + content" approach whereby the tables and fields (structure) as well as the vocabulary (content) are standardized, allowing for integration of data across multiple systems while maintaining data integrity. The model also supports a wide range of data types, including electronic health records (EHR), claims data, and patient registries.

The OMOP CDM is maintained by the Observational Health Data Science and Informatics (OHDSI) community, an open science effort that aims to improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care[25]. The open-source nature of OHDSI allows for continuous community-driven improvements, making it adaptable to emerging healthcare needs[26–28].

Data Partner Calls

Any organization with access to a data source in Europe could apply to be included in the EHDEN network. In this context, a data source is defined as a distinct repository of healthcare-related data pertaining to a specific set of individuals. Except for the COVID-19 Rapid Collaboration Call, the 7 data partner calls executed between September 2019 and October 2022 were aligned to similar timelines for data partner identification, grant awarding, and initiation of data harmonization (figure 1). In each call, candidate partner organizations with access to one or more electronic healthcare databases applied to the EHDEN Harmonization Fund for a grant to implement or enhance their database (supplemental 2). Data Partners were selected based on 3 criteria: data impact (size, coverage, quality), network impact (track record, uniqueness within network) and readiness (willingness to participate, governance) (supplemental 1), reviewed by a Data Source Priori-

tization Committee (DSPC). Each application was reviewed and scored by two reviewers and the top applicants per round were awarded a grant.



Figure 1: Timeline of Data Call 7 (2022–2023) within the European Health Data & Evidence Network (EHDEN). The figure illustrates key milestones in the selection, contracting, and data harmonization process for Data Partners onboarded during this call.

Data Standardization

Once DPs were identified and grants awarded, each data source underwent standardization to the OMOP CDM. A crucial factor in EHDEN's long-term sustainability and success was the recruitment and training of local small-to-medium enterprises (SMEs). These SMEs were brought on board through separate calls from the EHDEN consortium and certified via the EHDEN Academy education program concluded by an on-site/online training. SMEs played a pivotal role in supporting data partners throughout the extract, transform, and load (ETL) process by providing guidance and expertise. In total 64 SMEs across 22 countries were certified by EHDEN to support DPs (see www.ehden.eu for more details about these SMEs).

The ETL process followed by the data partners and supported by the SMEs was largely uniform, as outlined by Voss et al., and involved four key steps: 1) summarizing the native data, 2) creating the ETL specification, 3) mapping source vocabulary codes, and 4) implementing the ETL[24]. This standardized approach ensured transparency in the followed procedure and adherence to the conventions in converting data sources to the CDM, while also allowing DPs to benefit from the SMEs' specialized knowledge.

To promote semantic interoperability across heterogeneous healthcare systems, all source codes for diagnoses, medications, procedures, and measurements were mapped to standardized vocabularies (e.g., SNOMED CT, RxNorm, LOINC) as required by the OMOP CDM. For example, the UK Biobank contributed data that included SNOMED CT-coded diagnoses from electronic health records alongside custom-coded fields for self-reported conditions and blood pressure measurements. Mapping these nonstandard elements involved a combination of automated matching and manual curation followed by expert

validation. This process facilitated consistent interpretation of clinical concepts across countries and enhanced the analytical interoperability of the network[29].

Payments were structured based on output; to receive full funding DPs were required to meet 3 different milestones. The ETL specification document entitled data partners to 30%, ETL implementation and infrastructure released the next 40%, and the final 30% was received by the data partner after final inspection of the harmonized data (supplemental 3).

Data Quality

Each milestone was reviewed by an EHDEN consortium member who was part of the Milestone review committee. The ETL specification document required by milestone 1 was evaluated to ensure the mapping adhered to the OMOP CDM conventions and that the data partners/SMEs had a good understanding of the CDM and their own native data[23,30]. Milestone 2, the ETL implementation, had multiple review steps. The infrastructure was investigated to be sure the data partners were using a supported database platform[31]. The vocabulary mapping was evaluated to ensure most, if not all, source codes were included. The Data Quality Dashboard (DQD) was developed by EHDEN Work Package 5 to provide a standard structure for quality assurance[32]. It was utilized by data partners throughout the ETL process to continually improve the standardized data sources[33]. As described by Voss, et al the median number of times a data partner ran the DQD was 3 with an interquartile range of 2-7. In general, conformance issues to the OMOP CDM were identified and addressed in initial runs with more complex site-specific vocabulary mapping issues were addressed in subsequent runs. Data partners used the default failure thresholds in the first run. These were updated to reflect the nuances of each data source in later runs of the software[24]. In milestone 3, final DQD results as well as the CDM Inspection Report were reviewed[34]. Once approved, the data partner then entered their information into the EHDEN portal, an online platform open to the public designed to catalog metadata on each data source [35].

Analyses

An individual data source was considered one entry in the EHDEN portal. Data sources were categorized based on country, person count, the levels of care represented (primary, secondary or mixed), why a person was included, and how data were captured. These categories were ascertained from the data source description and metadata provided to the EHDEN portal and verified with the data partners.

There are three reasons persons could be included in an EHDEN data source (person inclusion), as defined by: **population**, where a person enters the data source because they live in a certain geographical location or because they are registered with a practice or insurer; **encounter**, where a person enters the data source upon a visit to a healthcare provider, for any medical reason; or **disease**, where a person enters the data source when satisfying specific criteria, i.e. a person has a specific medical condition. The most restrictive reason was chosen as the classification for each data source.

We identified which types of data capture methods each source contained, it could be one or more of the following: electronic health record (EHR), a bill or adjudicated claim record for health services rendered (claim), measurements taken and results recorded (lab), a set of required information collected about participants in a registry (case report form), patient reported data (survey), documents analyzed by pulling structured data from unstructured data using a natural language processing algorithm (NLP), or death information from an official source or government entity (death certificate). If the data source did not provide this information they were categorized as unknown.

RESULTS

As of September 1, 2024, there are 210 data sources in the EHDEN portal. Figure 2 shows all countries and the number of data sources available in each. The data sources span 30 countries with the largest representation from Italy, Great Britain, and Spain with 13.0%, 12.5% and 11.5% of the total data sources in the network, respectively. The mean number of persons per data source is 2,147,161 and the median number of persons is 457,664.

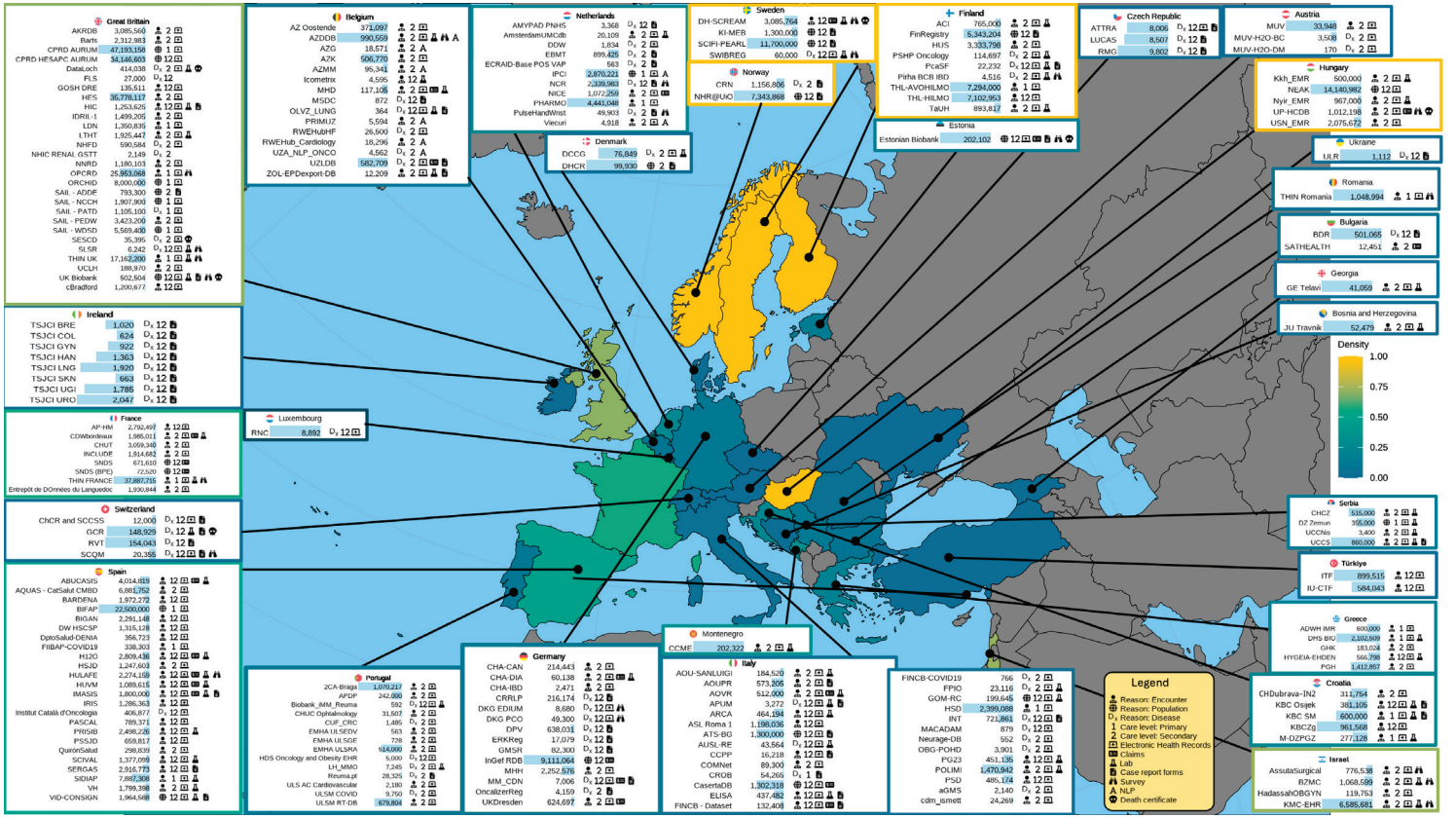


Figure 2: Geographic distribution and characteristics of data sources included in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The map displays country-level data density based on the number of data sources relative to national population size. Overlaid symbols represent key metadata for each source, including total person count, care setting (primary, secondary, or mixed), data capture methods (e.g., electronic health records, laboratory, or claims), and the reason for person inclusion (e.g., encounter-based, disease-specific, or population-based).

Table 1 provides the complete list of data sources and their attributes. One row in the table equates to one data source. The first column lists the data partner, which is the name of the institution or organization that is the custodian of the data source. The individual data sources are identified by an acronym which is also how they are identified in the EHDEN portal. Country of origin is represented by the two-digit country code. The number of persons, the person inclusion method, and care level are also provided. Each data capture category has its own column in the table. If a data source uses one of the capture methods, that box is colored green in the table.

Looking at care setting, 46.7% (98) of data sources represent data from the secondary setting only while 42.4% (89) represent data from mixed settings (primary and secondary). A comparatively smaller set of 11.0% (23) represent data only from the primary care setting (Table 2). Looking at the ways in which persons are included in the data sources, 55.7% (117) do so through healthcare encounters, 32.9% (69) through disease-specific data collection, and 11.4% (24) through population-based sources.

Figure 3 shows the number of data sources that receive information through each capture method and each combination of capture methods. EHR is the most common, with 74.7% (157) of data sources reporting at least one capture method as EHR. Over half of those data sources (85) report EHR as their only method for receiving data. Lab is the second most common way data sources capture information as it is reported in 29.5% (62) of data sources. Unlike EHR, lab data is more likely to be coupled with another data capture method as only one data source lists lab as the singular way they receive information.

Table 2: Stratification of data sources in the European Health Data & Evidence Network (EHDEN) by method of person inclusion and care level. Person inclusion reflects the basis by which individuals are represented in the database—healthcare encounters, disease-specific inclusion, or population-based inclusion. Care settings indicate whether data were captured in primary care, secondary (hospital) care, or across both (mixed).

Care Level	***Person Inclusion***			***Total, %***
	Disease	*Encounter*	*Population*	
Mixed	40	34	15	89 (42.4%)
Primary	2	14	7	23 (11%)
Secondary	27	69	2	98 (46.7%)
Total, %	69 (32.9%)	117 (55.7%)	24 (11.4%)	210

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network.

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Centro Clínico Académico – Braga, Associação (2CA-Braga)	2CA-Braga	PT	1070217	Encounter	Secondary	■							
INCLIVA	ABUCASIS	ES	4014819	Encounter	Mixed	■	■	■					
The wellbeing services county of Southwest Finland, VarHa	ACI	FI	765000	Encounter	Secondary	■		■					
Innovative Medical Research SA	ADWH IMR	GR	600000	Encounter	Primary	■							
Fondazione Casa Sollievo della Sofferenza	aGMS	IT	2140	Disease	Secondary	■							
Akrivia Health	AKRDB	GB	3085560	Encounter	Secondary	■							
Amsterdam UMC	Amsterda-mUMCdb	NL	20109	Encounter	Secondary	■		■					
Stichting VUmc	AMYPAD PNHS	NL	3368	Disease	Mixed				■				
AZIENDA OSPEDALIERO UNIVERSITARIA SAN LUIGI GONZAGA	AOU-SANLUIGI	IT	184520	Encounter	Secondary	■		■					
University Hospital of Parma	AOUPR	IT	573205	Encounter	Secondary	■			■				
Azienda Ospedaliera Universitaria Integrata Verona	AOVR	IT	512000	Encounter	Secondary	■	■	■					
Assistance Publique - Hopitaux de Marseille	AP-HM	FR	2792497	Encounter	Mixed	■							
APDP	APDP	PT	242000	Encounter	Secondary	■							
Azienda Ospedaliero-Universitaria di Modena	APUM	IT	3272	Disease	Mixed	■		■	■				

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Servei Català de la Salut	AQUAS - CatSalut CMBD	ES	6881752	Encounter	Secondary	■							
FONDAZIONE TOSCANA GABRIELE MONASTERIO PER LA RICERCA MEDICA E DI SANITA PUBBLICA (FTGM)	ARCA	IT	464194	Encounter	Mixed	■		■					
ASL Roma 1	ASL Roma 1	IT	1198036	Encounter	Mixed	■							
Assuta medical centers	AssutaSurgical	IL	776538	Encounter	Secondary	■				■			
ATS Bergamo	ATS-BG	IT	1300000	Population	Mixed	■			■				
Institute of Rheumatology	ATTRA	CZ	8006	Disease	Mixed	■			■				
Marco Massari (IRCSSE)	AUSL-RE	IT	43564	Disease	Mixed	■		■					
Az Oostende	AZ Oostende	BE	371097	Encounter	Secondary	■							
AZ Delta	AZDDB	BE	990559	Encounter	Secondary	■		■		■	■		
VZW AZ Groeninge	AZG	BE	18571	Encounter	Secondary	■					■		
AZ Klina	AZK	BE	506770	Encounter	Secondary	■							
AZ Maria Middelares	AZMM	BE	95341	Encounter	Secondary						■		
Servicio Navarro de Salud Osasunbidea (SNS-O)	BARDENA	ES	1972272	Encounter	Mixed	■							
Barts Health NHS Trust	Barts	GB	2312983	Encounter	Secondary	■							
National Scientific Programme "E-Health in Bulgaria"	BDR	BG	501065	Disease	Mixed	■			■				

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Agencia Española de Medicamentos y Productos Sanitarios, AEMPS	BIFAP	ES	22500000	Population	Primary	■							
Instituto Aragonés de Ciencias de la Salud (IACS)	BIGAN	ES	2291148	Encounter	Mixed	■							
Instituto de Medicina Molecular	Biobank_iMM_Reuma	PT	592	Disease	Mixed	■		■					
Bnai Zion Medical Research Foundation and Infrastructure Development Health Services	BZMC	IL	1068599	Encounter	Secondary	■		■		■			
Inspire-srl	CasertaDB	IT	1302318	Population	Mixed	■	■						
Connected Bradford	cBradford	GB	1200677	Encounter	Mixed	■							
Clinical Center of Montenegro	CCME	ME	202322	Encounter	Secondary	■		■					
Casa di Cura Privata del Policlinico (CCPP)	CCPP	IT	16218	Encounter	Mixed	■			■				
ISMETT	cdm_ismett	IT	24269	Encounter	Secondary	■							
Bordeaux University Hospital	CDWbordeaux	FR	1985011	Encounter	Secondary	■	■	■					
Charité - Universitätsmedizin	CHA-CAN	DE	214443	Encounter	Secondary	■							
Charité - Universitätsmedizin	CHA-DIA	DE	60138	Encounter	Secondary	■	■	■					
Charité - Universitätsmedizin	CHA-IBD	DE	2471	Encounter	Secondary	■							

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Institute of Social and Preventive Medicine, University of Bern	ChCR and SCCSS	CH	12000	Disease	Mixed	■			■				
Clinical-hospital center Zvezdara	CHCZ	RS	515000	Encounter	Secondary	■		■					
Clinical Hospital Dubrava	CHDubrava–IN2	HR	311754	Encounter	Secondary	■							
Centro Hospitalar Universitário de Coimbra (CHUC)	CHUC Ophtalmology	PT	31507	Encounter	Secondary	■							
Centre Hospitalier Universitaire de Toulouse	CHUT	FR	3059340	Encounter	Secondary	■							
Modena Oncology Center - Azienda Ospedaliera Modena	COMNet	IT	89300	Encounter	Secondary	■							
Clinical Practice Research Datalink (CPRD)	CPRD AURUM	GB	47193158	Population	Primary	■							
Clinical Practice Research Datalink (CPRD)	CPRD HESAPC AURUM	GB	34146603	Population	Mixed	■							
The Norwegian Cancer Registry	CRN	NO	1156806	Disease	Secondary	■			■				
Basilicata Cancer Registry	CROB	IT	54265	Disease	Primary				■				
Krebsregister Rheinland-Pfalz	CRRLP	DE	216174	Disease	Mixed				■				
CUF	CUF_CRC	PT	1485	Disease	Secondary	■							
DataLoch	DataLoch	GB	414038	Disease	Secondary	■		■				■	
Center for Surgical Science (CSS)	DCCG	DK	76849	Disease	Secondary	■		■					

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Amsterdam UMC	DDW	NL	1834	Disease	Secondary	■							
Stockholm CREAtinine Measurements Project	DH-SCREAM	SE	3085764	Encounter	Mixed		■	■	■	■		■	
University of Southern Denmark	DHCR	DK	99930	Population	Secondary				■				
DIGITAL HEALTH SOLUTIONS SA	DHS BIO	GR	2102509	Encounter	Primary	■		■					
German Cancer Society (DKG)	DKG EDIUM	DE	8680	Disease	Mixed	■				■			
German Cancer Society (DKG)	DKG PCO	DE	49300	Disease	Mixed	■				■			
Hospital de Denia	DptoSalud-DENIA	ES	356723	Encounter	Mixed	■							
University of Ulm, ZIBMT	DPV	DE	638031	Disease	Mixed				■				
Research Institute - Hospital de la Santa Creu i Sant Pau	DW HSCSP	ES	1315128	Encounter	Mixed	■		■					
Primary Healthcare Center Zemun	DZ Zemun	RS	355000	Population	Primary	■		■					
EBMT: The European Society for Blood and Marrow Transplantation	EBMT	NL	899425	Disease	Secondary				■				
European Clinical Research Alliance on Infectious Diseases (ECRAID)	ECRAID-Base POS VAP	NL	563	Disease	Secondary				■				
Centre Hospitalier Universitaire de Montpellier	eDOL Entrepôt de DOnnées du Languedoc	FR	1930844	Encounter	Secondary	■							
Fondazione IRCCS Policlinico San Matteo	ELISA	IT	437482	Encounter	Mixed	■		■	■				

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
EGAS MONIZ HEALTH ALLIANCE	EMHA ULSEDV	PT	563	Encounter	Secondary	■							
EGAS MONIZ HEALTH ALLIANCE	EMHA ULSGE	PT	728	Encounter	Secondary	■							
EGAS MONIZ HEALTH ALLIANCE	EMHA ULSRA	PT	514000	Encounter	Secondary	■							
European Rare Kidney Disease Registry (ERKReg)	ERKReg	DE	17079	Disease	Mixed				■				
University of Tartu	Estonian Biobank	EE	202102	Population	Mixed	■	■		■	■		■	
FIIBAP	FIIBAP-COVID19	ES	338303	Encounter	Primary	■							
Fondazione IRCCS Istituto Neurologico Carlo Besta	FINCB - Dataset	IT	132408	Encounter	Mixed	■	■		■				
Fondazione IRCCS Istituto Neurologico Carlo Besta FINCB	FINCB-COVID19	IT	766	Disease	Secondary	■							
FinRegistry (Institute of Molecular Medicine Finland (FIMM), University of Helsinki)	FinRegistry	FI	5343204	Population	Mixed				■				
Queen Mary University of London	FLS	GB	27000	Disease	Mixed								■
Fondazione Poliambulanza Istituto Ospedaliero	FPIO	IT	23116	Disease	Secondary	■		■					
Geneva Cancer Registry	GCR	CH	148929	Disease	Mixed			■	■			■	
Telavi Regional Hospital	GE Telavi	GE	41059	Encounter	Secondary	■		■					
GENERAL HOSPITAL OF KAVALA	GHK	GR	183024	Encounter	Secondary	■							

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
MS Forschungs- und Projektentwicklungs-GmbH	GMSR	DE	82300	Disease	Mixed	■			■				
Grande Ospedale Metropolitano "Bianchi-Melacrino-Morelli"	GOM-RC	IT	199645	Population	Mixed	■		■					
GOSH	GOSH DRE	GB	135511	Encounter	Mixed	■							
Fundacion de Investigacion Biomedica del Hospital Universitario 12 de Octubre	H12O	ES	2809436	Encounter	Mixed	■	■	■	■				
Hadassah OBGYN	HadassahOBGYN	IL	119753	Encounter	Secondary	■							
Hospital Distrital de Santarém (HDS)	HDS Oncology and Obesity EHR	PT	5000	Disease	Mixed	■							
Harvey Walsh Ltd	HES	GB	35778117	Encounter	Secondary	■							
Health Informatics Centre (HIC)	HIC	GB	1253625	Encounter	Mixed	■		■	■				
SIMG	HSD	IT	2399088	Encounter	Primary	■							
Hospital Sant Joan de Déu	HSJD	ES	1247603	Encounter	Secondary	■							
Fundación para la Investigación del Hospital Universitario La Fe de la Comunidad Valenciana (HULAFE)	HULAFE	ES	2274159	Encounter	Mixed	■	■	■	■	■			
Hospital District of Helsinki and Uusimaa	HUS	FI	3333798	Encounter	Secondary	■							
Virgen Macarena University Hospital	HUVM	ES	1089615	Encounter	Mixed	■	■	■	■				

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
DIAGNOSTIC & THERAPEUTIC CENTER OF ATHENS "HYGEIA" SINGLE MEMBER SOCIETE ANONYME	HYGEIA-EHDEN	GR	566798	Encounter	Mixed	■		■					
Icometrix	Icometrix	BE	4595	Encounter	Mixed			■					
Lancashire and South Cumbria Integrated Care Board	IDRIL-1	GB	1499205	Encounter	Secondary	■							
Fundacio Institut d'Investigacions Mèdiques (FIMIM)	IMASIS	ES	1800000	Encounter	Mixed	■			■				
Lille University Hospital	INCLUDE	FR	1914682	Encounter	Secondary	■							
InGef - Institute for Applied Health Research Berlin GmbH	InGef RDB	DE	9111064	Population	Mixed		■						
Institut Català d'Oncologia	Institut Català d'Oncologia	ES	406877	Disease	Mixed	■							
Fondazione Istituto Nazionale dei Tumori	INT	IT	721861	Disease	Mixed	■			■				
NO GRANT	IPCI	NL	2870221	Population	Primary	■		■			■		
Consorci Corporació Sanitària Parc Taulí	IRIS	ES	1286363	Encounter	Mixed	■							
Istanbul University	ITF	TR	899515	Encounter	Mixed	■							
IUC Cerrahpaşa TIP Fakületesi	IU-CTF	TR	584043	Encounter	Mixed	■							
E-MEDIT D.O.O. & Hospital Travnik	JU Travnik	BA	52479	Encounter	Secondary	■		■					

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
IN2 d.o.o. & Clinical Hospital Center Osijek	KBC Osijek	HR	381105	Encounter	Mixed	■		■	■				
IGEA d.o.o. & University Hospital Center Sestre milosrdnice	KBC SM	HR	600000	Encounter	Primary	■		■	■				
Hierarchia & University Hospital Centre Zagreb	KBCZg	HR	961568	Encounter	Mixed	■							
MEB KI	KI-MEB	SE	1300000	Population	Mixed				■				
Bács-Kiskun Megyei Kórház a Szegedi Tudományegyetem Általános Orvostudományi Kar Oktató Kórháza	Kkh_EMR	HU	500000	Encounter	Secondary	■		■					
The Directorate of Government Medical Centers at the Israeli Ministry Of Health	KMC-EHR	IL	6585681	Encounter	Secondary	■		■		■			
Lambeth DataNet	LDN	GB	1350835	Encounter	Primary	■							
Hospital da Luz Learning Health	LH_MMO	PT	7245	Disease	Secondary	■	■	■				■	
Leeds Teaching Hospitals	LTHT	GB	1925447	Encounter	Secondary	■		■					
OAKS Consulting s.r.o.	LUCAS	CZ	8507	Disease	Mixed				■				
MCS Grupa d.o.o. & Health Care Center of Primorje-Gorski Kotar County	M-DZPGZ	HR	277128	Encounter	Primary	■		■					

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Azienda Ospedaliera SS Antonio e Biagio e Cesare Arrigo	MACADAM	IT	879	Disease	Mixed	■							
Medaman	MHD	BE	117105	Encounter	Secondary	■	■	■					
Hanover Medical School	MHH	DE	2252576	Encounter	Secondary	■							
CancerDataNet GmbH	MM_CDN	DE	7006	Disease	Mixed	■	■		■				
University MS Center	MSDC	BE	872	Disease	Mixed				■				
Medical University of Vienna	MUV	AT	33948	Encounter	Secondary	■							
Medical University of Vienna	MUV-H2O-BC	AT	3508	Disease	Secondary	■							
Medical University of Vienna	MUV-H2O-DM	AT	170	Disease	Secondary	■							
IKNL	NCR	NL	2339983	Disease	Mixed			■	■	■			
National Institute of Health Insurance Fund Management Hungary	NEAK	HU	14140982	Population	Mixed	■							
AO Card. G. Panico - Center for Neurodegenerative Diseases and Aging Brain	Neurage-DB	IT	552	Disease	Secondary	■							
Queen Mary University of London	NHFD	GB	590584	Disease	Secondary	■							
King's College London	NHIC RENAL GSTT	GB	2149	Disease	Secondary			■					■
University of Oslo	NHR@UiO	NO	7343868	Population	Mixed				■				
National Intensive Care Evaluation foundation	NICE	NL	1072259	Encounter	Secondary	■	■						

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
UK National Neonatal Research Database	NNRD	GB	1180103	Encounter	Secondary	■							
Szabolcs-Szatmár-Bereg Megyei Kórházak és Egyetemi Oktatókórház	Nyir_EMR	HU	967000	Encounter	Secondary	■		■					
Bambino Gesù Children's Hospital	OBG-POHD	IT	3901	Disease	Secondary	■							
Onze-Lieve-Vrouwziekenhuis Aalst-Asse-Ninove	OLVZ_LUNG	BE	364	Disease	Mixed	■		■	■				
GermanOncology	OncalizerReg	DE	4159	Disease	Secondary				■				
Optimum Patient Care Limited	OPCRD	GB	25953068	Encounter	Primary	■				■			
Royal College of General Practitioners (RCGP)	ORCHID	GB	8000000	Population	Primary	■							
Rioja Salud	PASCAL	ES	789371	Encounter	Mixed	■							
University of Turku (Prostate Cancer Registry of South West Finland)	PcaSF	FI	22232	Disease	Mixed	■		■	■				
ASST Papa Giovanni XXIII	PG23	IT	451135	Encounter	Mixed	■		■					
Papageorgiou General Hospital	PGH	GR	1412857	Encounter	Secondary	■							
STIZON	PHARMO	NL	4441048	Encounter	Primary	■							
BCB Medical Ltd	Pirha BCB IBD	FI	4516	Disease	Secondary	■		■		■			
Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico	POLIMI	IT	1470942	Encounter	Secondary	■		■					

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
UZ Brussel	PRIMUZ	BE	5594	Encounter	Secondary			■			■		
Fundació Institut d´Investigació Sanitària Illes Balears	PRISIB	ES	2498226	Encounter	Mixed	■		■					
IRCCS Policlinico San Donato	PSD	IT	485174	Encounter	Mixed	■		■					
Finnish Clinical Biobank Tampere	PSHP Oncology	FI	114697	Disease	Secondary	■		■					
Parc Sanitari Sant Joan de Déu	PSSJD	ES	659817	Encounter	Mixed	■							
Harm Slijper	PulseHandWrist	NL	49903	Disease	Secondary				■	■			
Quironsalud	QuirónSalud	ES	298839	Encounter	Secondary	■							
Registo Portugues de Doentes Reumaticos	Reuma.pt	PT	28325	Disease	Secondary				■				
Czech Myeloma Group	RMG	CZ	9802	Disease	Mixed				■				
Registre National du Cancer du Luxembourg	RNC	LU	8892	Disease	Mixed	■							
Vaud Cancer Registry	RVT	CH	154043	Disease	Mixed				■				
LynxCare	RWEHub_Cardiology	BE	18296	Encounter	Secondary						■		
LynxCare	RWEHubHF	BE	26500	Disease	Secondary	■							
SAIL Databank	SAIL - ADDE	GB	793300	Population	Secondary				■				
SAIL Databank	SAIL - NCCH	GB	1907900	Population	Primary	■							
SAIL Databank	SAIL - PATD	GB	1105100	Disease	Primary	■							

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
SAIL Databank	SAIL - PEDW	GB	3423200	Encounter	Secondary	■							
SAIL Databank	SAIL - WDSD	GB	5569400	Population	Primary	■		■				■	
SAT Health	SATHEALTH	BG	12451	Encounter	Secondary		■						
Gothenburg University	SCIFI-PEARL	SE	11700000	Population	Mixed			■	■				
Servicio Cántabro de Salud and IDIVAL	SCIVAL	ES	1377099	Encounter	Mixed	■		■					
HUG and SCQM	SCQM	CH	20355	Disease	Mixed	■			■	■			
Consellería de Sanidade	SERGAS	ES	2916773	Encounter	Mixed	■			■				
University of Edinburgh	SESCD	GB	35395	Disease	Secondary	■						■	
SIDIAP - The Information System for Reseach in Primary Care	SIDIAP	ES	7887308	Encounter	Primary	■		■					
King's College London	SLSR	GB	6242	Disease	Mixed	■		■		■			
Health Data Hub	SNDS	FR	671610*	Population	Mixed		■						
Bordeaux PharmacoEpi	SNDS (BPE)	FR	72520*	Population	Mixed		■						
SWIBREG	SWIBREG	SE	60000	Disease	Mixed	■		■		■			
Pirkanmaa Hospital District	TaUH	FI	893817	Encounter	Secondary	■		■					
CEGEDIM HEALTH DATA	THIN FRANCE	FR	37887715	Encounter	Primary	■		■		■			
CEGEDIM HEALTH DATA	THIN Romania	RO	1048994	Encounter	Primary	■				■			
CEGEDIM HEALTH DATA	THIN UK	GB	17162200	Encounter	Primary	■		■		■			

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
Finnish Institute of Health and Welfare	THL-AVOHILMO	FI	7294000	Encounter	Primary	■							
Finnish Institute for Health and Welfare (THL)	THL-HILMO	FI	7102953	Encounter	Mixed	■							
Trinity St James's Cancer Institute	TSJCI BRE	IE	1020	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI COL	IE	624	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI GYN	IE	922	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI HAN	IE	1363	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI LNG	IE	1920	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI SKN	IE	663	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI UGI	IE	1785	Disease	Mixed				■	■			
Trinity St James's Cancer Institute	TSJCI URO	IE	2047	Disease	Mixed				■	■			
Clinical centre of Nis	UCCNis	RS	3400	Encounter	Secondary	■		■					
Clinical Center of Serbia	UCCS	RS	860000	Encounter	Secondary	■		■	■	■			
University College London Hospitals	UCLH	GB	188970	Encounter	Secondary	■							
University College London (UCL) (UK Biobank)	UK Biobank	GB	502504	Population	Mixed	■		■	■	■		■	
University Medicine Dresden	UKDresden	DE	624697	Encounter	Secondary	■	■						
National Cancer Institute	ULR	UA	1112	Disease	Mixed				■				

Table 1: Overview of 210 standardized real-world data sources in the European Health Data & Evidence Network (EHDEN) as of September 1, 2024. The table includes the full list of data sources by country, total number of persons represented, person inclusion, care level, and data capture methods. These attributes provide essential context for understanding the scope and scale of data available for real-world evidence generation within the EHDEN network. (*continued*)

Data Partner	Data Source Acronym	Country	Person Count	Person Inclusion	Care Level	EHR	Claim	Lab	Case Report Form	Survey	NLP	Death Certificate	Unknown
ULS AC Cardiovascular	ULS AC Cardiovascular	PT	2180	Encounter	Secondary	■							
ULSM	ULSM COVID	PT	9750	Disease	Secondary	■							
Unidade Local de Saúde de Matosinhos	ULSM RT-DB	PT	679804	Encounter	Secondary	■							
University of Pécs	UP-HCDB	HU	1012198	Encounter	Secondary	■	■			■		■	
Semmelweis University	USN_EMR	HU	2075672	Encounter	Secondary	■							
University Hospital Antwerp	UZA_NLP_ONCO	BE	4562	Disease	Secondary						■		
Universitaire Ziekenhuizen KU Leuven	UZLDB	BE	582709	Disease	Secondary	■	■		■				
Vall d'Hebrón Hospital Campus	VH	ES	1799398	Encounter	Secondary	■		■					
FISABIO-HSRU	VID-CONSIGN	ES	1964588	Population	Mixed	■		■	■				
VieCuri Medisch Centrum	Viecuri	NL	4918	Encounter	Secondary	■					■		
Ziekenhuis Oost-Limburg	ZOL-EPDexport-DB	BE	12209	Encounter	Secondary	■		■	■				

*This is a subset of the full data source.

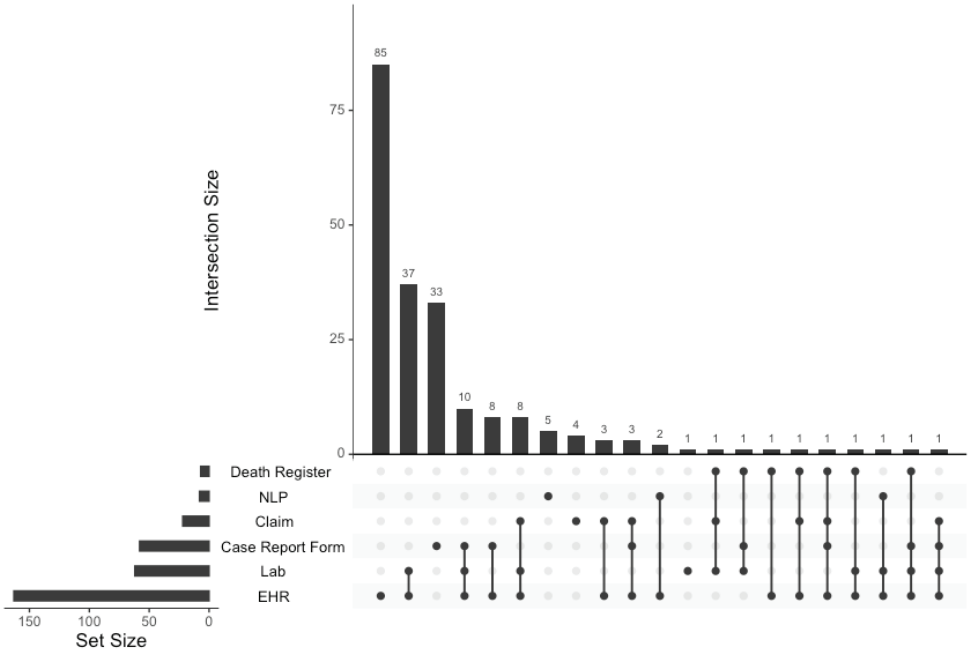


Figure 3: The frequency and overlap of different data capture methods used across 210 standardized real-world data sources in the European Health Data and Evidence Network (EHDEN) as of September 1, 2024.

DISCUSSION

The varied healthcare data across Europe, as demonstrated by the summary of 210 data sources in EHDEN from 30 countries, underscores the critical need to generate evidence from more than one data source to comprehensively represent the healthcare needs or experiences of the entire European population. Across the person inclusion and care levels represented in the network, the data sources are well distributed, emphasizing how healthcare systems, populations, and data capture methods can differ substantially. While 74.7% of the data sources report EHR as at least one of their data capture methods, only 40.4% report EHR as their only data capture method. The other 34.3% report some combination of EHR, Lab, Case Report Form, Claim, NLP, and Death Register data, showcasing the tremendous heterogeneity of data available in Europe.

Prior Initiatives

Prior initiatives like EU-ADR and IMI-EMIF laid the groundwork for EHDEN, with learnings from those projects directly impacting this project[15,18,36,37]. EU-ADR demonstrated the feasibility of building a federated data network for large-scale drug safety monitoring in Europe using common data analysis files. IMI-EMIF made the first transition from using

Common Input Files like those in EU-ADR to the OMOP CDM but it was not scalable due to the lack of funds and need for trained SMEs, both problems which EHDEN addressed.

Sustainability and Success of EHDEN

The sustainability of the EHDEN initiative has been achieved through a combination of mechanisms that foster shared leadership, collaboration, and long-term value creation. One key factor has been the stimulation and enablement of both national and European collaborations. The establishment of OHDSI National Nodes has provided a platform for data partners within individual countries to collaborate, share best practices, and enhance data quality[38]. These nodes facilitate national-level harmonization while ensuring compliance with local regulations and coding systems, thereby strengthening the network's integrity. Beyond this, EHDEN's adoption in multiple European projects has further expanded its influence, including its pivotal role in enabling large-scale initiatives such as the Data Analysis and Real World Interrogation Network (DARWIN EU®). This has also influenced how the European Federation of Pharmaceutical Industries and Associations (EFPIA) is standardizing its data, demonstrating EHDEN's impact across sectors.

EHDEN has also delivered economic value by creating local ecosystems that support SMEs and data partners. Through the Harmonisation Fund, EHDEN has injected resources into the European healthcare data landscape, with the return on investment yielding a multiplier effect. By recruiting and training SMEs through the EHDEN Academy, the initiative has built local expertise to support data partners throughout the ETL process, ensuring decentralized and sustainable support for the network.

One of the goals of EHDEN has been to standardize health data—akin to utilities like electricity or the internet—essential and accessible to a rapidly growing number of stakeholders across Europe. Now that EHDEN has transitioned from a project under IHI to the non-profit EHDEN Foundation, the focus has shifted to sustain, expand, and improve the network while leveraging the harmonized data for evidence generation. This next phase aims to generate meaningful real-world evidence (RWE) for research and regulatory purposes. A recent report by The European Commission on the future of European competitiveness highlights EHDEN's foundational role in shaping the future of the European Health Data Space, further solidifying its legacy as a critical driver of innovation and collaboration in European healthcare[39].

The success of EHDEN in harmonizing data to the OMOP CDM has led to significant advances in methods research and evidence generation[40–43]. Many of the data partners involved in EHDEN have used their standardized data to conduct analyses across a broad spectrum of use cases. For instance, several studies have been conducted to describe the natural history of diseases, the safety and effectiveness of treatments, and healthcare utilization patterns across diverse populations[44–46]. One clear example is a multina-

tional network cohort study by Li et al., which used evidence generated from EHDEN data partners to characterize background incidence rates of adverse events of special interest related to COVID-19 vaccines[45]. EHDEN's standardized data has also enabled and improved the development of predictive models, allowing for personalized predictions of treatment outcomes and disease progression[47,48]. The harmonization of data has facilitated large-scale population-level studies, which are crucial for understanding trends in public health and informing healthcare policy decisions[49,50]. These examples of evidence generation illustrate the broad applicability of the data in the network, which serves both academic researchers and regulatory agencies. A regularly updated list of EHDEN-supported studies and publications is maintained on the EHDEN website (https://zenodo.org/communities/ehden/records?q=&l=list&p=1&s=10&sort=newest), providing a comprehensive overview of the diverse applications of the network in regulatory, clinical, and methodological research.

A notable demonstration of EHDEN's success is the network's use in providing timely information on medicines under surveillance due to shortages in multiple European countries. More than fifty Data Partners contributed data to a study titled "Incidence, Prevalence, and Characterization of Medicines with Suggested Drug Shortages in Europe[51]." This study represents the largest observational database study conducted across Europe, both in terms of the number of databases involved and its geographic scope. The findings will support European efforts to monitor the use of critical medicines, contributing to the global fight against medicine shortages.

The EHDEN data network has also affected other European collaboratives around RWD. IMI projects like PIONEER, BigData@Heart, EU-PEARL, HARMONY, also use the OMOP CDM and have partly continued the mapping work done in EHDEN[52–55]. In the EMA commissioned Data Analysis and Real World Interrogation Network (DARWIN EU®) initiative, among the 20 DPs onboarded in the first two years, 16 are also EHDEN DPs[56].

Future Directions

Building on the progress achieved through the EHDEN network, several key areas offer opportunities for future development. One priority is fostering sustained engagement with data partners. Continuous collaboration will be essential to ensure that data partners remain active contributors to the network by regularly updating and improving their data contributions. Strategies to incentivize engagement, provide ongoing support, and ensure mutual value will be vital for the network's long-term success, particularly as efforts shift toward more robust evidence generation and ongoing enhancements in data quality.

Expanding the network's reach and optimizing its databases for specific research use cases are also key areas for growth. With 210 data sources currently included, there is a

significant opportunity to onboard additional data partners and expand the network's coverage across Europe. Future studies will also help identify gaps where further data optimization is required, such as refining mappings or addressing specific quality issues to ensure that the evidence generated is robust, reproducible, and generalizable.

Finally, the newly established EHDEN Foundation will play a critical role in these efforts. By securing funding and fostering collaborations, the Foundation can drive the onboarding of new data partners, address emerging research questions, and ensure that EHDEN continues to adapt to the evolving healthcare landscape. It also serves as a point of entry for external researchers, who may engage with the network and propose studies through its federated framework. These directions will position the network to remain a cornerstone for real-world evidence generation in Europe, supporting both research and regulatory innovation.

CONCLUSION

The results of this study demonstrate that the identification, harmonization, and standardization of data sources through EHDEN have contributed significantly to understanding the diverse RWD landscape and advancement of evidence generation across Europe. These efforts are not only improving observational health research but are also influencing broader regulatory initiatives, such as DARWIN EU®, which builds on the foundational work of EHDEN to leverage real-world data for regulatory decision-making. Now that the initiative has transitioned to the EHDEN Foundation, there is an opportunity to focus even more on generating high-quality evidence, further solidifying the role of real-world data in improving healthcare and informing policy decisions across Europe.

ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to all EHDEN Data Partners who contributed their time, effort, and expertise in harmonizing their data to the OMOP Common Data Model. Their commitment to data quality and transparency has been fundamental to the success of this initiative.

We extend our appreciation to the Small-to-Medium Enterprises (SMEs) that played a pivotal role in supporting Data Partners through the extract, transform, and load (ETL) process. Their expertise in data mapping and technical implementation has been invaluable in ensuring high-quality and standardized data across the network.

We would also like to acknowledge the contributions of the EHDEN Work Packages and their leads, whose dedication to data harmonization, quality assessment, infrastructure development, training, and sustainability efforts have made this project possible. Their leadership and vision have shaped EHDEN into a robust and scalable federated data network.

Finally, we extend our thanks to all members of the EHDEN Consortium who have worked tirelessly to build and sustain this network. This includes researchers, data scientists, software engineers, governance and regulatory experts, and the broader Observational Health Data Sciences and Informatics (OHDSI) community, whose open-science collaboration and innovation continue to drive the success of EHDEN.

We recognize that the continued success of EHDEN is the result of collective contributions from numerous stakeholders across Europe, and we sincerely appreciate the efforts of everyone involved.

ABBREVIATIONS

EHDEN	European Health Data & Evidence Network
RWD	Real-world data
OMOP CDM	Observational Medical Outcomes Partnership Common Data Model
RWE	Real-world evidence
SME	Small-to-medium enterprise
ETL	Extract, transform, and load
DQD	Data Quality Dashboard
RCT	Randomized controlled trial
IMI	Innovative Medicines Initiative
IHI	Innovative Health Initiative
DPs	Data Partners
EHR	Electronic health records
DSPC	Data Source Prioritization Committee
DARWIN EU®	Data Analysis and Real World Interrogation Network
EFPIA	European Federation of Pharmaceutical Industries and Associations

DECLARATIONS

Author Contributions

All authors (CB, MJS, MM, EAV, MC, PRR, PBR) were involved in data collection. CB, MJS, PBR, MM, PRR were involved in the study design, analysis, and interpretation of results. CB, MJS, PBR, PRR contributed to writing and all authors revised and approved the final draft.

Funding

This project has received support from the European Health Data and Evidence Network (EHDEN) project. EHDEN received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Competing Interests

CB, MJS, EAV, PBR, are employees of Johnson & Johnson and hold stock and stock options. PRR works for a department that receives/received unconditional research grants from Amgen, Chiesi, Johnson and Johnson, UCB Biopharma, the European Medicines Agency and the Innovative Medicines Initiative.

Patient Consent

Not required

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Data Availability

All data supporting this work can be found on the EHDEN portal at https://portal.ehden.eu/summary.

SUPPLEMENTARY MATERIAL

Multimedia Appendix 1: EHDEN Data Partner call description.
Multimedia Appendix 2: EHDEN sub-grant agreement model. Av
Multimedia Appendix 3: EHDEN framework for quality benchmarking.

All supplementary material is available online at https://www.jmir.org/2025/1/e74119.

REFERENCES

1. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018.
2. Heads of Medicines Agency, European Medicines Agency. HMA-EMA Joint Big Data Taskforce Summary Report. 2019 Feb 13;48.
3. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology. Available from: https://encepp.europa.eu/encepp-toolkit/methodological-guide_en [accessed Sep 6, 2024]
4. Jonker C, Bakker E, Kurz X, Plueschke K. Contribution of patient registries to regulatory decision making on rare diseases medicinal products in Europe. Frontiers in Pharmacology 2022; doi: 10.3389/FPHAR.2022.924648
5. Eskola SM, Leufkens H, Bate A, Bruin ML, Gardarsdottir H. The Role of Real-World Data and Evidence in Oncology Medicines Approved in EU in 2018-2019. Journal of Cancer Policy 2023; doi: 10.1016/J.JCPO.2023.100424
6. Bolislis WR, Fay M, Kühler TC. Use of Real-world Data for New Drug Applications and Line Extensions. Clinical Therapeutics 2020; doi: 10.1016/J.CLINTHERA.2020.03.006
7. European Medicines Agency. Guidance on good pharmacovigilance practices (GVP). Module VIII. Post-authorisation safety studies. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-viii-post-authorisation-safety-studies-rev-3_en.pdf [accessed Aug 6, 2024]
8. Santoro A, Genov G, Spooner A, Raine J, Arlett P. Promoting and Protecting Public Health: How the European Union Pharmacovigilance System Works. Drug Saf 2017 Oct;40(10):855–869. PMID:28735357
9. Goedecke T, Morales DR, Pacurariu A, Kurz X. Measuring the impact of medicines regulatory interventions - Systematic review and methodological considerations. Br J Clin Pharmacol 2018 Mar;84(3):419–433. PMID:29105853
10. OECD. Health at a Glance: Europe 2022: State of Health in the EU Cycle. Paris: Organisation for Economic Co-operation and Development; 2022. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2022_507433b0-en [accessed Aug 22, 2024]
11. Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clinical Pharmacology & Therapeutics 2019;106(1):36–39. doi: 10.1002/cpt.1426
12. Bakker E, Plueschke K, Jonker C, Kurz X, Starokozhko V, Mol P. Contribution of Real-World Evidence in European Medicines Agency's Regulatory Decision Making. Clinical pharmacology and therapy 2022; doi: 10.1002/CPT.2766
13. Flynn RWV, Plueschke K, Quinten C, Strassmann V, Duijnhoven RG, Gordillo-Marañón M, Rueckbeil M, Cohet C, Kurz X. Marketing Authorization Applications Made to the European Medicines Agency in 2018–2019: What was the Contribution of Real-World Evidence? 2021; doi: 10.1002/CPT.2461
14. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, Cazzola W, Coloma P, Berni R, Diallo G, Oliveira JL, Avillach P, Trifirò G, Rijnbeek P, Bellentani M, van Der Lei J, Klazinga N, Sturkenboom M. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. eGEMs AcademyHealth; 2016; Available from: https://dash.harvard.edu/handle/1/26318777 [accessed Sep 6, 2024]
15. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Corrao G, Pedersen L, van der Lei J, Sturkenboom M, Consortium on behalf of the E-A.

Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiology and Drug Safety 2011;20(1):1–11. doi: 10.1002/pds.2053
16. Trifirò G, Coloma P, Rijnbeek P, Romio S, Mosseveld B, Weibel D, Bonhoeffer J, Schuemie M, Lei J, Sturkenboom M. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? Journal of internal medicine 2014; doi: 10.1111/JOIM.12159
17. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010 Nov 2;153(9):600–606. PMID:21041580
18. Reynolds RF, Kurz X, de Groot MCH, Schlienger RG, Grimaldi-Bensouda L, Tcherny-Lessenot S, Klungel OH. The IMI PROTECT project: purpose, organizational structure, and procedures. Pharmacoepidemiol Drug Saf 2016 Mar;25 Suppl 1:5–10. PMID:27038353
19. EFPIA Recommendations on a Connected Data System in Europe. Available from: https://www.efpia.eu/news-events/the-efpia-view/efpia-news/efpia-recommendations-on-a-connected-data-system-in-europe/ [accessed Sep 6, 2024]
20. European Health Data Evidence Network. ehden.eu. Available from: https://www.ehden.eu/ [accessed Mar 12, 2021]
21. Innovative Health Initiative | IHI Innovative Health Initiative. 2024. Available from: http://www.ihi.europa.eu/ [accessed Sep 12, 2024]
22. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, Londhe A, Zhu V, Ryan PB. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553–564. PMID:25670757
23. Observational Health Data Sciences and Informatics. OMOP Common Data Model (CDM). Available from: https://ohdsi.github.io/CommonDataModel/
24. Voss EA, Blacketer C, van Sandijk S, Moinat M, Kallfelz M, van Speybroeck M, Prieto-Alhambra D, Schuemie MJ, Rijnbeek PR. European Health Data & Evidence Network-learnings from building out a standardized international health data network. J Am Med Inform Assoc 2023 Dec 22;31(1):209–219. PMID:37952118
25. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y-C, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574–578. PMID:26262116
26. Schuemie M, Reps J, Black A, Defalco F, Evans L, Fridgeirsson E, Gilbert JP, Knoll C, Lavallee M, Rao GA, Rijnbeek P, Sadowski K, Sena A, Swerdel J, Williams RD, Suchard M. Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research. Stud Health Technol Inform 2024 Jan 25;310:966–970. PMID:38269952
27. Chapter 4 The Common Data Model. The Book of OHDSI. Available from: https://ohdsi.github.io/TheBookOfOhdsi/
28. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, Dymshyts D, Hripcsak G. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association 2024 Mar 1;31(3):583–590. doi: 10.1093/jamia/ocad247
29. Papez V, Moinat M, Voss EA, Bazakou S, Van Winzum A, Peviani A, Payralbe S, Lara EG, Kallfelz M, Asselbergs FW, Prieto-Alhambra D, Dobson RJB, Denaxas S. Transforming and evaluating

the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. J Am Med Inform Assoc 2022 Oct 13;30(1):103–111. PMID:36227072
30. THEMIS | THEMIS Convention Repository for the OMOP CDM. Available from: https://ohdsi.github.io/Themis/ [accessed Sep 13, 2024]
31. OHDSI/DatabaseConnector. Observational Health Data Sciences and Informatics; 2025. Available from: https://github.com/OHDSI/DatabaseConnector [accessed May 7, 2025]
32. 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; doi: 10.1093/jamia/ocab132
33. Blacketer C, Voss EA, DeFalco F, Hughes N, Schuemie MJ, Moinat M, Rijnbeek PR. Using the Data Quality Dashboard to Improve the EHDEN Network. Applied Sciences Multidisciplinary Digital Publishing Institute; 2021 Jan;11(24):11920. doi: 10.3390/app112411920
34. EHDEN/CdmInspection: R Package to support quality control inspection of an OMOP-CDM instance. Available from: https://github.com/EHDEN/CdmInspection [accessed Sep 13, 2024]
35. EHDEN Portal. Available from: https://portal.ehden.eu/wordpress/index.php/landing-page-2/ [accessed Sep 13, 2024]
36. Klungel O, Kurz X, Groot MC, Schlienger R, Tcherny-Lessenot S, Grimaldi L, Ibañez L, Groenwold R, Reynolds R. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiology and drug safety 2016; doi: 10.1002/PDS.3968
37. Lovestone S. The European medical information framework: A novel ecosystem for sharing healthcare data across Europe. Learn Health Syst 2019 Dec 25;4(2):e10214. PMID:32313838
38. Observational Health Data Sciences and Informatics. National Nodes. Available from: https://www.ohdsi-europe.org/index.php/national-nodes [accessed Sep 27, 2024]
39. Mario Draghi. The Future of European Competitiveness. 2024. Available from: https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en#paragraph_47059
40. Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. Journal of Big Data 2024 Jan 3;11(1):7. doi: 10.1186/s40537-023-00857-7
41. Fridgeirsson EA, Williams R, Rijnbeek P, Suchard MA, Reps JM. Comparing penalization methods for linear models on large observational health data. J Am Med Inform Assoc 2024 Jun 20;31(7):1514–1521. PMID:38767857
42. Williams RD, den Otter S, Reps JM, Rijnbeek PR. The DELPHI Library: Improving Model Validation, Transparency and Dissemination Through a Centralised Library of Prediction Models. Stud Health Technol Inform 2023 May 18;302:139–140. PMID:37203630
43. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021 Jan;113:103655. PMID:33309898
44. Morales DR, Ostropolets A, Lai L, Sena A, DuVall S, Suchard M, Verhamme K, Rijnbeek P, Posada J, Ahmed W, Alshammary T, Alghoul H, Alser O, Areia C, Blacketer C, Burn E, Casajust P, Chan You S, Dawoud D, Golozar A, Gong M, Jonnagaddala J, Lynch K, Matheny M, Minty E, Nyberg F, Uribe A, Recalde M, Reich C, Scheumie M, Shah K, Shah N, Schilling L, Vizcaya D, Zhang L, Hripcsak G, Ryan P, Prieto-Alhambra D, Durate-Salles T, Kostka K. Characteristics and outcomes of COVID-19 patients with and without asthma from the United States, South Korea, and Europe. Journal of Asthma 2022 Feb 11; doi: 10.1080/02770903.2021.2025392
45. Li X, Ostropolets A, Makadia R, Shoaibi A, Rao G, Sena AG, Martinez-Hernandez E, Delmestri A, Verhamme K, Rijnbeek PR, Duarte-Salles T, Suchard MA, Ryan PB, Hripcsak G, Prieto-Alhambra

D. Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. BMJ British Medical Journal Publishing Group; 2021 Jun 14;373:n1435. doi: 10.1136/bmj.n1435

46. Voss EA, Shoaibi A, Yin Hui Lai L, Blacketer C, Alshammari T, Makadia R, Haynes K, Sena AG, Rao G, van Sandijk S, Fraboulet C, Boyer L, Le Carrour T, Horban S, Morales DR, Martínez Roldán J, Ramírez-Anguita JM, Mayer MA, de Wilde M, John LH, Duarte-Salles T, Roel E, Pistillo A, Kolde R, Maljković F, Denaxas S, Papez V, Kahn MG, Natarajan K, Reich C, Secora A, Minty EP, Shah NH, Posada JD, Garcia Morales MT, Bosca D, Cadenas Juanino H, Diaz Holgado A, Pedrera Jiménez M, Serrano Balazote P, García Barrio N, Şen S, Üresin AY, Erdogan B, Belmans L, Byttebier G, Malbrain MLNG, Dedman DJ, Cuccu Z, Vashisht R, Butte AJ, Patel A, Dahm L, Han C, Bu F, Arshad F, Ostropolets A, Nyberg F, Hripcsak G, Suchard MA, Prieto-Alhambra D, Rijnbeek PR, Schuemie MJ, Ryan PB. Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 databases: a multinational retrospective cohort study. EClinicalMedicine 2023 Apr;58:101932. PMID:37034358

47. Williams RD, Reps JM, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. Knee Surgery, Sports Traumatology, Arthroscopy 2021 Dec 6; doi: 10.1007/s00167-021-06799-y

48. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, Ryan PB, Park RW, Lim H-S, Rijnbeek PR. Feasibility and evaluation of a large-scale external validation approach for patientlevel prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. 2020 May 6; doi: 10.1186/s12874-020-00991-3

49. Lane JCE, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, Alghoul H, Alser O, Alshammari TM, Areia C, Biedermann P, Banda JM, Burn E, Casajust P, Fister K, Hardin J, Hester L, Hripcsak G, Kaas-Hansen BS, Khosla S, Kolovos S, Lynch KE, Makadia R, Mehta PP, Morales DR, Morgan-Stewart H, Mosseveld M, Newby D, Nyberg F, Ostropolets A, Park RW, Prats-Uribe A, Rao GA, Reich C, Rijnbeek P, Sena AG, Shoaibi A, Spotnitz M, Vignesh S, Suchard MA, Vizcaya D, Wen H, Wilde M de, Xie J, You SC, Zhang L, Lovestone S, Ryan P, Prieto-Alhambra D, consortium O-C-19. Risk of depression, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multinational network cohort study. Rheumatology 2020 Dec 25;(00):1–13. doi: 10.1093/rheumatology/keaa771

50. Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, He Y, Robinson DE, Pinedo-Villanueva R, Kolovos S, Duarte-Salles T, Sproviero W, Yu D, Speybroeck MV, Williams R, John LH, Hughes N, Sena AG, Costello R, Birlie B, Culliford D, O'Leary C, Morgan H, Burkard T, Prieto-Alhambra D, Ryan P. Opioid use, postoperative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study. 2019 Nov 7; doi: 10.1016/S2665-9913(19)30075-X

51. Pineda-Moncusí M, Rekkas A, Pérez ÁM, Leis A, Gomez CL, Bruninx E, Maljković F, Rodeiro J, Franz M, Mayer M-A, Eleangovan N, Natsiavas P, Şen S, Cooper S, Reisberg S, Manlik K, Sánchez-Sáez F, Pino B del, Uribe AP, Üresin AY, Bastić AD, Rodrigues AM, Palomar-Cros A, Verbiest A, Erdoğan B, Keuthage CD-, Torre CO, Beukelaar C de, Eteve-Pitsaer C, Gonçalves CF, Palma C de, Gavina C, Dedman D, Price DB, Balan DG, Enders D, Henke E, Scheurwegs E, Callewaert E, Martínez EP, Tan EH, Fey E, Prasser F, Staelens F, Nyberg F, Candore G, Mestdach G, Shachaf H, Li H, Reinecke I, López-Sánchez I, Bertolo J de la C, Evers J, Firmino-Machado J, Wastesson J, Bermúdez JLC, Ramírez-Anguita JM, Porkka K, Johnell K, Cool L, Kiss LZ, Moscetti L, Merkelbach M, Canelas-Pais M, Dominici M, Szilcz M, Puntoni M, Mosseveld M, Tadrous M, Bové M, Rappoport N, Barrio NG, Ettala O, Baili P, Pulido PP, Mayo PR, Prinsen P, Norris R,

Claire R, Lillini R, Lazzarelli S, Duarte-Salles T, Taveira-Gomes T, Jansen T, Keilholz U, Li X, Prieto-Alhambra D, Rijnbeek PR, Burkard T. Trends of drug use with suggested shortages and their alternatives across 41 real world data sources and 18 countries in Europe and North America. medRxiv; 2024. p. 2024.08.28.24312695. doi: 10.1101/2024.08.28.24312695

52. PIONEER - European Network of Excellence for Big Data in Prostate Cancer. PIONEER - European Network of Excellence for Big Data in Prostate Cancer. Available from: https://prostate-pioneer.eu/ [accessed Nov 7, 2024]
53. BigData@Heart > Home. Available from: https://www.bigdata-heart.eu/ [accessed Nov 7, 2024]
54. HARMONY is part of the IMI "Big Data for Better Outcomes" program. Available from: https://www.harmony-alliance.eu/en/harmony-is-part-of-the-imi-big-data-for-better-outcomes-program [accessed Nov 7, 2024]
55. EU-PEARL - Innovative Patient Centric Clinical Trial Platforms. EU-PEARL. Available from: https://eu-pearl.eu/ [accessed Nov 7, 2024]
56. European Medicines Agency. Darwin EU. Available from: https://www.darwin-eu.org/ [accessed Sep 26, 2024]

6

Validation of the Database Diagnostics method: Assessing Fitness-for-Purpose in Federated Networks

Clair Blacketer, MPH[1,2,4]
Patrick B. Ryan, PhD[1,3,4]
Marc A. Suchard, MD, PhD[1,5]
Mui Van Zandt[6],
Jason C. Hsu, PhD[7,8,9,10]
Thanh-Phuc Phan, PhD[7]
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
[6]Real World Solutions, IQVIA, Durham, NC, US
[7]International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan
[8]Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan
[9]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan
[10]Research Center of Data Science on Health Care Industry, College of Management, Taipei

Under Submission

ABSTRACT

Background

The rise of the secondary use of healthcare data and the emergence of federated data networks (FDNs) have significantly enhanced capabilities for conducting large-scale, multi-institutional clinical research. The success of these networks relies on the ability to quickly and efficiently select data sources best fit to answer specific research questions without jeopardizing patient privacy. In this study, we introduce and validate the Database Diagnostics method developed to address this need.

Objective

Using the Large-scale Evidence Generation and Evaluation across a Network of Databases for Type 2 Diabetes Mellitus (LEGEND-T2DM) study as a reference case we validate the Database Diagnostics method, demonstrating it's potential as a novel, scalable approach to fitness-for-purpose assessments in federated networks.

Methods

Database Diagnostics is a rule-based method that compares study-specific requirements to summary statistics collected from databases in federated networks to determine those with adequate representation that may be fit to answer the research question. To validate this approach, the requirements for the LEGEND-T2DM study were fed into Database Diagnostics. The results for both the exposure and outcome phenotypes were then compared to the current published results of LEGEND-T2DM to assess how well the method identified databases that were ultimately able to generate evidence for the study.

Results

Across the 12 databases with full Database Diagnostics and LEGEND-T2DM results, Database Diagnostics did well identifying databases with any representation of the drugs of interest, but tended to overestimate the number of databases with at least 1000 persons exposed. This was due to the method being unable to replicate the complex target cohorts using marginal summary statistics alone. Database Diagnostics performed very well in identifying the databases best fit to support individual outcomes, demonstrating high accuracy with a sensitivity of 100% and specificity of 92%. In addition, the method pinpointed why individual databases could not support certain outcomes, providing critical insight not available otherwise.

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.

INTRODUCTION

The rise of the secondary use of healthcare data and the emergence of federated data networks (FDNs) have significantly enhanced capabilities for conducting large-scale, multi-institutional clinical research. Defined here as networks in which data remain locally housed at each institution but can be queried and analyzed through standardized, distributed methods, FDNs enable researchers to generate evidence across multiple institutions without centralizing patient-level information, effectively addressing privacy and governance constraints associated with such data [1–3]. Successful examples such as the Observational Health Data Sciences and Informatics (OHDSI) network, the European Health Data & Evidence Network (EHDEN) and the Data Analysis and Real World Interrogation Network (DARWIN EU®) illustrate how federated analyses can be operationalized at scale to support comparative effectiveness research, regulatory needs, and other real-world evidence generation efforts [4–8].

However, a critical gap remains in efficiently identifying whether databases within FDNs are fit-for-purpose, or having the necessary elements required for answering a study question. This is related to but separate from ensuring those elements are of high quality. Data quality as a general property of a database has been addressed in prior publications, but systematically evaluating whether databases contain the necessary data elements *to support specific study questions* remains a significant challenge [9–14]. This gap can hinder researchers' ability to plan effective, interpretable, and reproducible analyses in federated settings.

Our approach to addressing this challenge leverages data standardization, which is central to the success of federated research networks and often achieved through the use of common data models (CDMs). The Observational Medical Outcomes Partnership (OMOP) Common Data Model, in particular, enables structural and semantic harmonization of data, allowing analytic tools to be developed once and executed consistently across all network partners [15–17]. With adoption by hundreds of data partners internationally, the OMOP CDM has become a foundational component of distributed real-world evidence initiatives [18].

Existing OHDSI tools such as Cohort Diagnostics and Study Diagnostics partially address the need for fit-for-purpose assessments by enabling the evaluation of phenotype definitions and study design elements across databases. Cohort Diagnostics allows researchers to inspect characteristics such as concept set (code set list) composition, demographic distribution, and temporal trends [19]. Study Diagnostics builds on this by providing prespecified, empirical checks on study design elements like cohort balance and model assumptions, ensuring robustness and reliability prior to full-scale analysis [20]. However, these approaches require running study-specific code on each database, which may be

impractical or inefficient at scale for the first task of identifying databases in a network qualified to answer specific study questions. Thus, there remains a need for scalable methods capable of rapidly assessing database fitness-for-purpose.

Recognizing the limitation of existing methods, we propose Database Diagnostics as a scalable alternative. Database Diagnostics systematically identifies databases potentially able to inform study questions using pre-computed summary statistics, eliminating the need for database-specific execution of study code during fit-for-purpose assessments. Leveraging standardized outputs generated during common database characterization workflows, such as metrics on person counts by concept (code), demographic distribution, temporal coverage, and domain availability this approach enables researchers to efficiently assess the fitness of databases prior to detailed study execution, reducing inefficiencies and increasing transparency in the planning stage. It is important to note, however, that the information used by this method is comprised of only marginal summary statistics, meaning no detailed information on combinations of variables is available. This restriction limits the potential for inappropriate use of the data and misinterpretation of results.

In this paper, we introduce and validate the Database Diagnostics method using the Large-scale Evidence Generation and Evaluation across a Network of Databases for Type 2 Diabetes Mellitus (LEGEND-T2DM) study as a reference case [21–23]. Our validation demonstrates that Database Diagnostics provides a scalable, privacy-preserving, and reproducible approach to evaluate database fitness-for-purpose, ultimately enabling more reliable and efficient evidence generation in federated research networks.

METHODS

Database Diagnostics

Database Diagnostics is a rule-based method developed to assess whether a given database is likely to be fit-for-purpose for a specific observational study, without requiring access to patient-level data. It supports these assessments within federated data networks by analyzing a standardized set of precomputed, population-level summary statistics from OMOP CDM–mapped databases.

To ensure consistency across diverse data sources, the OHDSI community standardizes clinical concepts through the OMOP CDM's standardized vocabulary. Source codes from administrative claims and electronic health records (e.g., ICD-10, CPT, Read, LOINC) are mapped to standard concepts that represent clinical ideas in a common representation. This semantic harmonization enables cross-database analyses using a unified set of concept identifiers, ensuring that exposure and outcome definitions can be applied con-

sistently regardless of the underlying coding system. Database Diagnostics builds on this foundational capability by assessing whether the relevant standard concepts required for a particular study design are sufficiently represented in each database.

The method relies on a set of independent summary statistics that cannot be connected back to person-level data. Here these statistics are referred to collectively as a Database Profile (DbProfile), which is computed for each database using the open-source DbDiagnostics R package[2]. These statistics include (full list available in Appendix 1):

- Demographic characteristics (e.g., age, sex, race, year of birth)
- Temporal coverage (e.g., number and duration of periods of observable time per person, earliest and latest date available in the data)
- Domain content (e.g., person and record count by drug concept, conditions, procedures, observations, measurements, and devices)
- Measurement completeness (e.g., value distributions)
- Cross-domain overlaps (e.g., number of patients with records in multiple data domains, like conditions and drugs)

Each participating site executes the DbDiagnostics package locally to generate a DbProfile for each database (a single site may have multiple databases). All DbProfiles for all databases are then combined into one table for ease of use (see Figure 1). Once generated, these summary statistics can be reused indefinitely to assess feasibility for any number of future studies—eliminating the need for repeated execution. The typical use case involves an FDN where summary statistics from multiple sites are combined centrally for fit-for-purpose assessments. However, the approach can also be used within a single site to facilitate rapid initial feasibility screening.

To implement the fit-for-purpose assessment, comparative effectiveness study requirements are typically specified in terms of a target (T), comparator (C), and outcome (O). The target and comparator define the populations or treatment groups under comparison, while the outcome refers to the endpoint or event of interested evaluated independently from the target and comparator. These T/C/O specifications are translated into structured representations, each encoding the clinical logic and data elements necessary to evaluate whether a database contains sufficient information to support the intended research question. Database Diagnostics applies a structured library of independent evaluation rules to each combination of settings object and database. Rules are grouped into three core categories:

2 https://github.com/ohdsi/DbDiagnostics

- **Concept Coverage**: Determines whether the required concepts are present in the database and at sufficient density.
- **Criteria Availability**: Assesses whether the database contains the clinical elements required to operationalize the study, such as demographic requirements, necessary data domains like procedures or measurements, or other inclusion/exclusion criteria. This group is broken into two subcategories:
    - **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.
    - **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.
- **Temporal Distribution**: Evaluates whether the database provides sufficient temporal coverage—either in terms of calendar time (e.g., availability of data from 2010–2015) or patient-level follow-up (e.g., number of persons with ≥365 days of observation).

The complete set of diagnostic rules and their operational definitions are provided in the DbDiagnostics R package.

For each database and diagnostic setting, the method produces a binary classification by comparing the results to pre-specified thresholds that can be changed by the user:

- **Pass**: all required elements are present and meet pre-specified thresholds.
- **Fail**: one or more critical rules are violated, indicating that cohort instantiation would be infeasible or likely to yield low-confidence results.

Each output is accompanied by metadata summarizing the rationale for the classification, including rule-level failures. Figure 1 shows the full process from DbProfile generation to feasibility assessment.

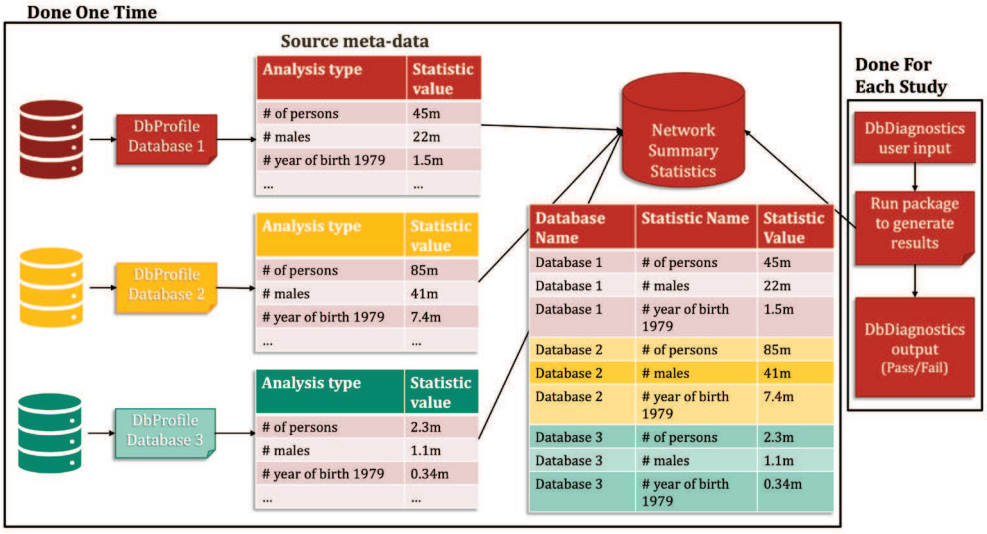


Figure 1: DbDiagnostics process, from summary statistic (DbProfile) generation to feasibility assessment (DbDiagnostics output). One-time processes are separate from processes conducted for each study.

An important constraint of this approach is that the evaluation rules rely on independent summary statistics derived separately from each database. Consequently, these independent metrics limit the ability to assess interactions or dependencies between elements directly, potentially impacting the comprehensiveness of the database suitability assessment.

Validation Using LEGEND-T2DM

To evaluate the accuracy of Database Diagnostics, 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) across a network of 36 databases. The study protocol, cohort definitions, and analytic code were developed and executed using OHDSI tools, with full code available online [https://ohdsi-studies.github.io/LegendT2dm/Protocol].

For the purposes of this validation, we used two publicly released LEGEND-T2DM cohort diagnostics results sets:

- **Drug Class Cohorts**: [https://data.ohdsi.org/LegendT2dmClassCohortExplorer/]
- **Outcome Cohorts**: [https://data.ohdsi.org/LegendT2dmOutcomeCohortExplorer/]

From these result sets we could extract the same summary statistics used by Database Diagnostics. Additionally, these dashboards report, for each cohort and each database in the LEGEND-T2DM study, whether the cohort was instantiated successfully and how many persons met the cohort criteria. We used these results as a reference standard for evaluating the performance of Database Diagnostics.

The validation proceeded in three phases:

Drug Class Exposure Assessment

We first evaluated whether Database Diagnostics correctly identified databases containing exposures to the relevant drug classes defined in the LEGEND-T2DM protocol. According to the LEGEND-T2DM cohort diagnostics, a database was considered to support a drug class if the corresponding exposure cohort included at least 1,000 individuals.

To compare these results with Database Diagnostics, we examined each database–cohort pair and determined:

1. Whether the database had any evidence of exposure to the specified drug class (i.e., any non-zero cohort count)
2. Whether the database met the ≥1,000 person threshold used in LEGEND-T2DM.

We then compared these empirical indicators to the corresponding Database Diagnostics classifications ("pass" or "fail") for each exposure cohort. This allowed us to assess how well Database Diagnostics anticipated both the presence and sufficient volume of drug class exposures necessary for inclusion in comparative analyses.

Outcome Support Assessment

Next, we evaluated whether Database Diagnostics could identify databases that were unable to support specific outcome definitions. Using the LEGEND-T2DM outcome cohort diagnostics, we flagged cases where an outcome cohort could not be instantiated or returned a zero count. We then compared these cases to Database Diagnostics results to determine whether the method had correctly classified the database as a "fail" for that outcome.

Population Stratification Assessment*:*

To align with the heterogeneity analyses planned in the LEGEND-T2DM study, we conducted an assessment to determine whether Database Diagnostics could anticipate differences in population structure that may affect the feasibility of subgroup analyses. The LEGEND-T2DM protocol includes stratified evaluations of treatment effects across clinically relevant subgroups defined by age, gender, race, cardiovascular risk, and renal impairment. These subgroup analyses aim to support patient-focused treatment recommendations by identifying potential treatment effect heterogeneity.

As a preliminary feasibility assessment, we used Database Diagnostics to evaluate the underlying demographic distributions of patients across databases—specifically focusing on age, gender, and race, which are routinely captured in the OMOP CDM and represented in the Database Diagnostics summary statistics. Rather than applying the full heterogeneity framework across all T/C/O combinations, we selected a single representative T/C/O combination to estimate whether each database was likely to support stratified analyses within the following subgroups:

- Age: 18–44, 45–64, ≥65 years at index
- Gender: men
- Race: African American or Black

This focused analysis allowed us to explore whether Database Diagnostics could provide early signals about the feasibility of heterogeneity assessments without introducing excessive analytic overhead.

Databases Included in the Study

A total of 36 databases generated and submitted a Database Profile to the OHDSI Coordinating Center in the spring of 2023 initially as part of a broader initiative to support simultaneous execution of four federated network studies. They agreed to participate in the current analysis with Database Diagnostics results available for 31 of these databases; five were excluded due to data quality issues or incomplete metadata. Among the remaining 31 databases, 10 participated in the LEGEND-T2DM study. These 10 had complete cohort diagnostics results available for both the exposure and outcome cohorts, allowing for full validation of the Database Diagnostics classifications against the LEGEND-T2DM reference. Table 1 details each of the 10 databases including population, number of patients, calendar time, and description.

Table 1: List of databases included in the validation of Database Diagnostics compared to LEGEND-T2DM cohort diagnostics results for exposures and outcomes.

Data source	Population	Patients	History	Data capture process and short description
Administrative claims				
Merative™ MarketScan® Commercial Claims and Encounters (CCAE)	Commercially insured, < 65 years	142M	2000 –	Adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) from large employers and health plans who provide private healthcare coverage to employees, their spouses and dependents.
Merative™ MarketScan® Medicare Supplemental Database (MDCR)	Commercially insured, 65++ years	10M	2000 –	Adjudicated health insurance claims of retirees with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service or capitated health plans.

Table 1: List of databases included in the validation of Database Diagnostics compared to LEGEND-T2DM cohort diagnostics results for exposures and outcomes. (*continued*)

Data source	Population	Patients	History	Data capture process and short description
Merative™ MarketScan® Multi-State Medicaid Database (MDCD)	Medicaid enrollees, racially diverse	26M	2006 –	Adjudicated health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims.
Optum's Clinformatics Extended Data Mart (Optum)	Commercially or Medicare insured	85M	2000 –	Inpatient and outpatient healthcare insurance claims.
Electronic health records (EHRs)				
Columbia University Irving Medical Center (CIUMC)	Academic medical center patients, racially diverse	6M	1989 –	General practice, specialists, and inpatient hospital services from the New York-Presbyterian hospital and affiliated academic physician practices in New York.
IQVIA Disease Analyzer Germany (DAG)	Germany, general	37M	1992 –	Collection from patient management software used by general practitioners and selected specialists to document patients' medical records within their office-based practice during a visit.
Optum's Electronic Health Records (OptumEHR)	US, general	94M	2006 –	Clinical information, prescriptions, lab results, vital signs, body measurements, diagnoses, and procedures derived from clinical notes using natural language processing.
Longitudinal Patient Database in France (FranceLPD)	France, general	17M	2012 -	Anonymized patient records collected from Patient Management software used by doctors during an office visit to document patients' clinical records.
Taipei Medical University Database (TMUDB)	Taipei, academic medical center patients	3M	2003 -	General practice, specialists, and inpatient hospital services from the Taiwan Medical University Hospital.
Longitudinal Patient Database in Australia (LPDAU)	AU, general	2M	2013 -	EMR source data contains medical history, allergies, immunizations, contacts and medical events related to them (test results, diagnoses, biometrics, prescriptions)

Database Diagnostics Scope and Execution

The LEGEND-T2DM protocol defines six pairwise comparisons across four drug classes: DPP4 inhibitors (DPP4i), GLP1 receptor agonists (GLP1RA), SGLT2 inhibitors (SGLT2i), and sulfonylureas (SU). Each pairwise comparison was assessed for the relative risk of 32 pre-specified outcome definitions, resulting in a total of:

6 pairwise comparisons × 32 outcome cohorts = 192 unique T/C/O settings

In total, the DbDiagnostics package was executed across 31 databases, with full LEGEND-T2DM results available in the 10 listed above. This yielded:

192 diagnostic settings × 10 databases with full results =
1,920 database–cohort evaluations

For the assessment focused on demographic stratification, a single representative T/C/O combination was selected, and Database Diagnostics was used to estimate whether sufficient sample sizes existed within key demographic subgroups—age, gender, and race—across participating databases. This yielded:

(3 age evaluations + 1 gender evaluation + 1 race evaluation) × 10 databases with full results =
50 database–demographic evaluations

For each evaluation, Database Diagnostics produced a binary classification of "pass" or "fail," supported by rule-level diagnostics describing which criteria were met or failed. These results formed the basis for comparison to the LEGEND-T2DM cohort diagnostics, enabling empirical evaluation of the method's validity.

For this study, we created 192 diagnostic settings objects derived from the LEGEND-T2DM protocol, comprising 6 pairwise comparisons and 32 outcome cohorts (see Validation section). All settings objects are available in JSON format in Appendix 2.

RESULTS

An example of the raw output is provided in Table 2. The FranceLPD database has records of both the DPP4I and GLP1RA drugs with counts of 129,190 and 148,300, respectively. The outcome 3-point MACE requires either a diagnosis either in the emergency room or inpatient setting and as FranceLPD contains only outpatient encounters it fails on those requirements.

The full raw output is available as a supplemental file. In some cases, the specification is given as CONCEPT_IDs, like the proportion of patients with ethnicity criteria. All concepts and their meanings can be accessed online at https://athena.ohdsi.org.

Table 2: Raw DbDiagnostics output for the DPP4I vs GLP1RA comparison for the outcome of 3-point MACE in the FranceLPD database.

Analysis Name	Database ID	Statistic	Value	Proportion	Specification	Threshold	Status
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with calendar time	17484250	1.000	201212-202306	0.05	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients in age range	14696600	0.841	estimated persons in the database aged 18 - 100	0.05	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with age at first observation	14389470	0.823	age 18 - age 100	0.01	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with at least 1 Emergency Room visit	0	0.000	Emergency Room visits desired	0	fail
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with ethnicity criteria	17484220	1.000	0	0.05	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with gender criteria	17353690	0.993	8507, 8532	0.05	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with at least 1 Inpatient visit	0	0.000	Inpatient visits desired	0	fail
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with longitudinal criteria	8817990	0.504	365 days	0.05	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with Race criteria	17484220	1.000	0	0.05	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with required comparator concepts	148300	0.008	GLP1RA main	0	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with records in required domain	12538550	0.717	condition, drug	0.05	pass

Table 2: Raw DbDiagnostics output for the DPP4I vs GLP1RA comparison for the outcome of 3-point MACE in the FranceLPD database. (*continued*)

Analysis Name	Database ID	Statistic	Value	Proportion	Specification	Threshold	Status
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with required outcome concepts	152550	0.009	3-point MACE	0	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	Proportion of patients with required target concepts	129190	0.007	DPP4I main	0	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	The estimated smallest sample size the study will yield	32083	0.002	> 1000	0	pass
DPP4I main_GLP1RA main_3-point MACE	FranceLPD	The estimated largest sample size the study will yield	129190	0.007	> 1000	1000	pass

Exposure Cohorts

As seen in figure 2, across the 31 databases, Database Diagnostics identified any representation of the concepts of interest for each target drug class in more than 90% of databases. Nearly all databases were predicted by the method to contain sufficient representation to support comparative effectiveness studies for DPP4I and SU, while for GLP1RA fewer passed the ≥1,000 exposed persons threshold (77.4%).

Among 10 databases with full LEGEND cohort diagnostics available, Database Diagnostics tended to overestimate the number of databases that could construct sufficiently sized exposure cohorts using the ≥1,000 persons threshold (figure 3). For example, 90.0% of databases were identified by Diagnostics as having ≥1,000 persons exposed to GLP1RA, but only 50.0% were confirmed by LEGEND. A similar discrepancy was seen for SGLT2I (100.0% vs. 60.0%).

Figure 4 shows a confusion matrix comparing the Database Diagnostics prediction of databases that meet the ≥1,000 exposed persons threshold to the LEGEND T2DM results. A true positive is determined by a Database Diagnostics prediction that a database contains ≥1,000 persons exposed to a given drug class and a result of ≥1,000 persons identified in the database after the full phenotype definition was generated. A true negative is determined by a prediction that a given database contains < 1,000 persons exposed to the drug class and a result of < 1,000 persons after the full phenotype definition was generated. A false positive is a prediction of ≥1,000 exposed persons when the full phenotype found < 1,000 persons and a false negative is a prediction of < 1,000 exposed persons when the full phenotype found ≥1,000 exposed. There were 30 true positives, 9 false positives, 1 true negative and 0 false negatives. This resulted in both a sensitivity and negative predictive value of 100.0%, specificity of 10.0%, and positive predictive value of 76.9%.

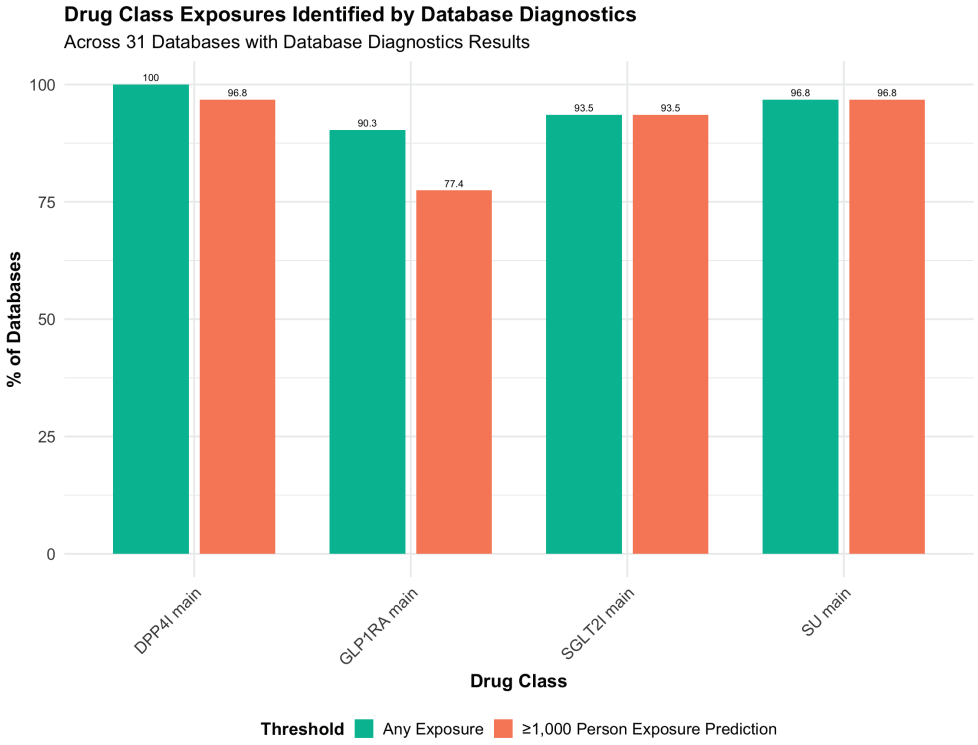


Figure 2: Comparison of drug class exposures identified in the 31 databases with results from Database Diagnostics.

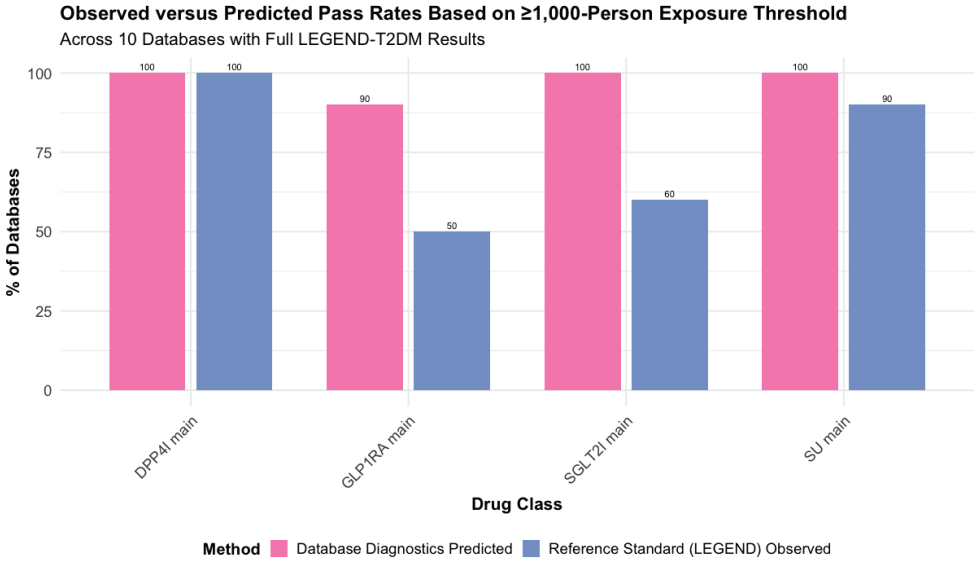


Figure 3: Observed versus predicted percent of databases with ≥1,000 persons exposed to each of four second-line anti-hyperglycemic drug classes across the 10 databases with full LEGEND results.

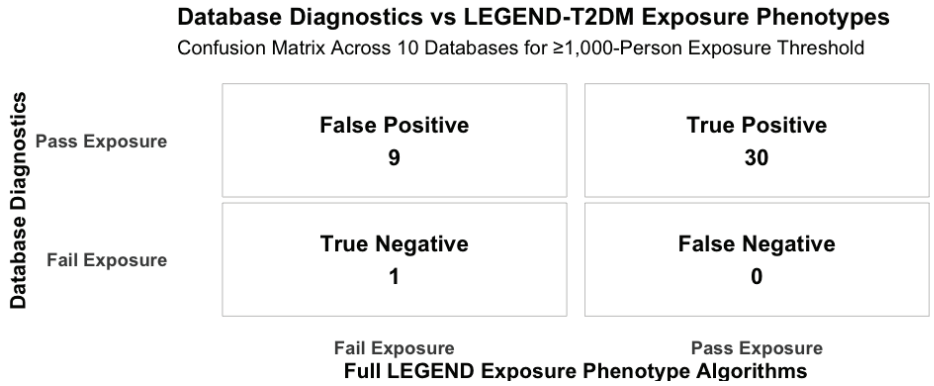


Figure 4: Confusion matrix showing the counts of false positive, true positive, true negative, and false negative predictions made by Database Diagnostics of the databases that meet the ≥1,000 persons threshold.

Outcome Cohorts

Figure 3 shows a confusion matrix comparing the Database Diagnostics prediction of whether a database can support an outcome to the actual Legend-T2DM phenotype result. A true positive is determined by a Database Diagnostics prediction of "Pass" for a given database x outcome and a result of >= 1 persons with the outcome in the database after the full phenotype definition was generated. Similarly, a true negative is determined by a prediction of "Fail" for a given database x outcome and a result of 0 persons with the outcome in the database after the full phenotype definition was generated. A false positive is a prediction of "Pass" and a result of 0 persons with the outcome after the generation of the phenotype. A false negative is a prediction of "Fail" and a result of >= 1 persons with the outcome once the full phenotype is generated. Overall, there were 272 true positives, 45 true negatives, 3 false positives, and 0 false negatives. This amounts to a positive predictive value of 98.9%, negative predictive value of 100%.

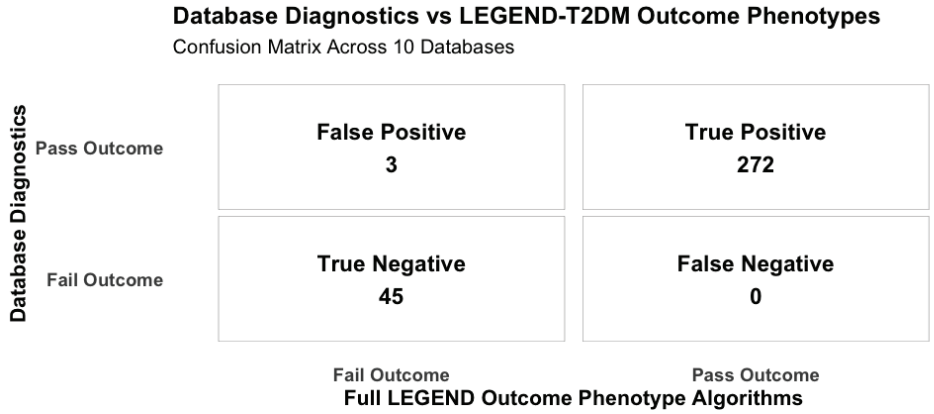


Figure 5: Confusion matrix showing the counts of false positive, true positive, true negative, and false negative predictions made by Database Diagnostics of outcome support by database.

Among the 45 true negatives, the most common reason for outcome failure (66.6%) was the absence of inpatient or emergency room visit data (figure 6). An additional 20.0% failed due to insufficient representation of the clinical concepts necessary to identify the outcome. In 4.4% of cases, the failure was attributable to the lack of death information, and another 4.4% of cases failed due to both incomplete concept representation and the absence of required visit types.

The false positives were identified incorrectly due to multiple reasons. In both the FranceLPD and LPDAU databases, the outcome Measured Renal Dysfunction was unable to run though Database Diagnostics predicted it would pass. This is because the outcome phenotype definition was missing required unit concept ids. The method does not currently evaluate measurement unit pairs, instead it only looks at the required measurement concepts, leading to this discrepancy. For the final false positive, Photosensitivity in the TMUCRD database, it was challenging to identify the issue. Database Diagnostics found approximately 30 persons with at least one record of any of the concepts used to define Photosensitivity, but the full phenotype definition returned zero persons. The hypothesis is that the records were in the incorrect domain table. Database Diagnostics found records with the concepts of interest in both the Condition_Occurrence table and the Observation table. It is possible that, at the time the full phenotype was run, that the records existed only in the Observation table and not the required Condition_Occurrence table, causing this difference.

When comparing the Database Diagnostics predictions against the full population stratification phenotypes, the method was able to identify 100% of databases with the ability to study the race, age, and gender subgroups.

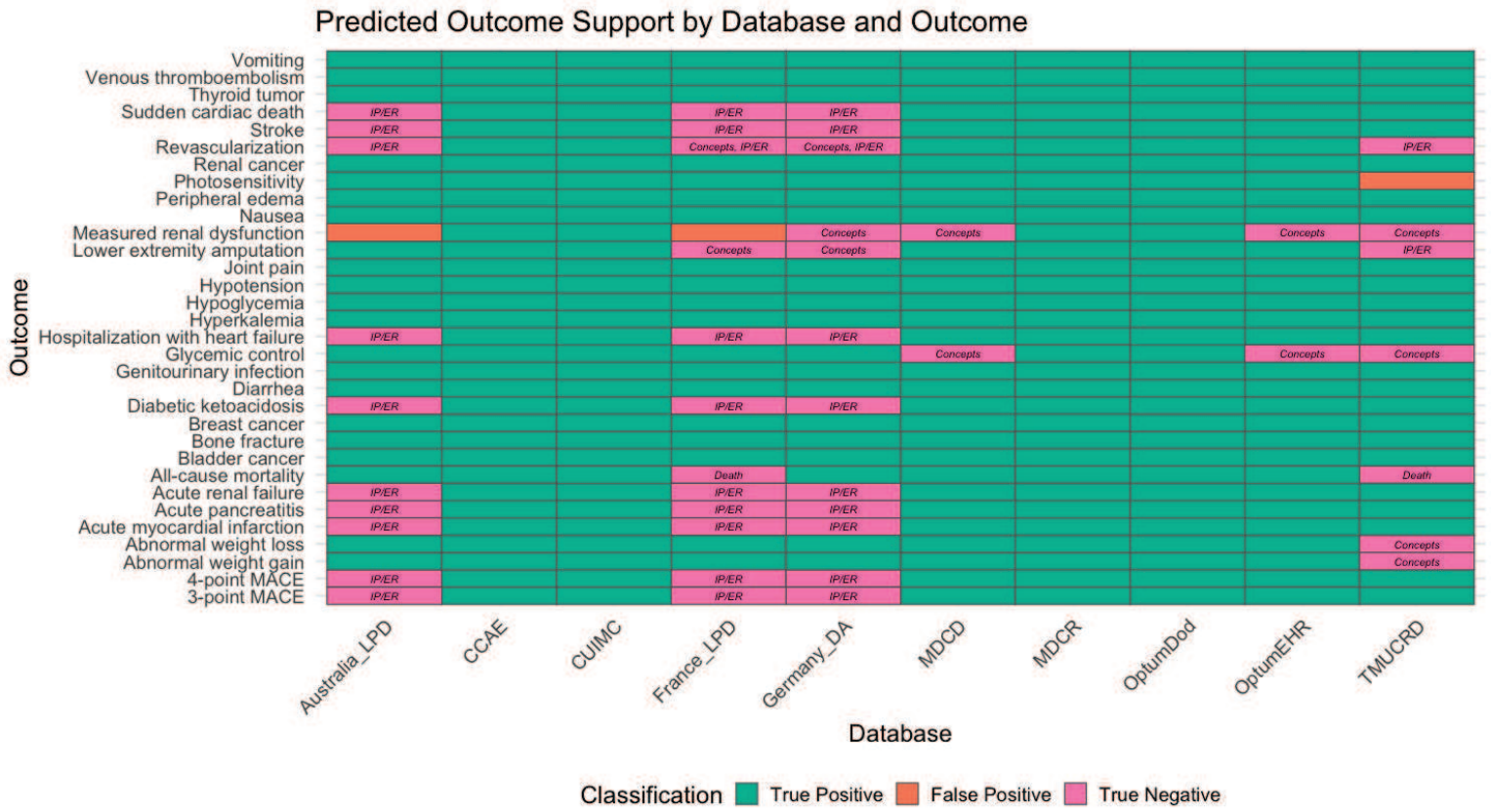


Figure 6: True positive, false positive and true negative outcomes by database and true negative reasons for phenotype failure.

DISCUSSION

Database Diagnostics is a novel, scalable approach to identifying whether databases in a federated network are fit-for-purpose for a given study question. In contrast to prior tools that evaluate phenotype logic or study validity after cohorts are instantiated, this method enables early fit-for-purpose assessment using only aggregate summary statistics. The distinction is critical: by relying solely on summary data, Database Diagnostics eliminates the need for patient-level access during study planning, substantially reducing the time, resources, and governance required to evaluate each database against the needs of a given analysis. In doing so, it fills a key methodological and operational gap in the planning phase of real-world evidence generation across distributed data networks.

Our validation against the LEGEND-T2DM study demonstrates that Database Diagnostics is especially effective at identifying databases that are unable to contribute to a given analysis. The method exhibited a negative predictive value of 100% when predicting outcome feasibility, suggesting that when the tool classifies a database as failing, it is highly reliable. This level of performance is valuable in federated settings where data access is limited and executing large-scale phenotype workflows across all sites can be time- and resource-intensive. By helping to avoid unnecessary execution in databases unlikely to yield results, the method supports efficient, transparent study planning.

The method also demonstrated strong performance in detecting the presence of relevant exposure and outcome concepts. However, it was less accurate in assessing whether a database contained a sufficiently large sample size to support population-level effect estimation. In several instances, Database Diagnostics overestimated database capability by identifying a drug class as sufficiently represented when subsequent phenotype instantiation found fewer than 1,000 eligible individuals. These false positives were especially common for less prevalent exposures such GLP1 receptor agonists and SGLT2 inhibitors. Notably, these discrepancies stemmed from the complex eligibility criteria used to define exposure cohorts in the study. Specifically, patients were required to be new users of one of the four drug classes, with no prior exposure to the others. Because Database Diagnostics relies on independent summary statistics, it cannot operationalize this type of longitudinal or exclusionary logic. While this limitation suggests opportunities to improve sample size estimation, the method's conservative approach of excluding only a small number of databases may be preferable to prematurely removing data sources that could ultimately generate evidence.

A key strength of Database Diagnostics lies in the transparency and granularity of its output. For each database-study combination, the method provides rule-level classifications that explicitly indicate which requirements were met and which were not. This level of detail enables users to pinpoint *why* a particular study or phenotype fails in a given database; information that is not readily available when the full study or phenotype is run. These diagnostics can uncover unexpected gaps in the data itself, such as missing domains or incomplete concept mappings, which may indicate issues with data standardization or the extract, transform, and load (ETL) process. At the same time, they can reveal problems in the study or phenotype design, such as missing unit concept IDs required to construct a valid measurement-based outcome, as seen in the Measured Renal Dysfunction example. By identifying both data-related and design-related issues before execution, Database Diagnostics serves as a powerful tool for iterative refinement and informed decision-making.

While Database Diagnostics demonstrated strong negative predictive performance, some false positives highlighted areas for improvement, particularly in refining rule thresholds. For example, in one database, the presence of a very small number of death records led the method to classify the All-Cause Mortality outcome as feasible. However, the number of deaths was far too low to support meaningful analysis, and the outcome was ultimately not instantiated. This case illustrates a key limitation: the method currently applies uniform thresholds that may not account for clinical or statistical context. In outcomes where a minimal number of events is essential to estimate effects with adequate precision, a more stringent threshold may be warranted. These findings suggest a need to recalibrate rule criteria to better differentiate between the mere presence of required elements and the presence of those elements at sufficient scale to support robust inference.

Looking ahead, Database Diagnostics could be extended through the development of more granular rules and additional criteria, such as evaluating measurement-unit pairs or adding age-group specific concept counts. Integrating this tool with other components of the OHDSI study design ecosystem such as Cohort Diagnostics and Study Diagnostics may enable a more comprehensive and automated pipeline for assessing fitness-for-purpose. Furthermore, although validated here using LEGEND-T2DM, the method is generalizable to any observational study executed across OMOP CDM standardized data sources, offering a reproducible and privacy-preserving approach to study planning across a broad range of clinical questions.

In summary, Database Diagnostics provides a practical, scalable, and privacy-preserving method for assessing whether a database is fit-for-purpose for a given study. By leveraging precomputed summary statistics and applying a structured rule-based approach, the method enables early, transparent evaluation of database feasibility without requiring patient-level data. Our validation using LEGEND-T2DM demonstrates its utility in identifying both feasible and infeasible study settings, with especially high reliability in flagging databases unlikely to support a given analysis. Beyond fit-for-purpose screening, the method offers unique diagnostic insights into *why* a particular study may fail, informing both study design and data curation efforts. While refinements are needed the core framework is robust and generalizable. As federated research networks continue to grow in scale and complexity, tools like Database Diagnostics will be essential to improving the efficiency, reproducibility, and transparency of real-world evidence generation.

APPENDIX 1

Full list of summary statistics collected by the DbDiagnostics R package

1. Number of persons
2. Number of persons by gender
3. Number of persons by year of birth
4. Number of persons by race
5. Number of persons by ethnicity
6. Number of persons by age, with age at first observation period
7. Number of persons by length of observation period, in 30d increments
8. Number of persons by observation period start month
9. Number of persons by observation period end month
10. Number of persons by number of observation periods
11. Number of persons with at least one day of observation in each month
12. Number of persons with at least one visit occurrence, by visit_concept_id
13. Number of persons with at least one condition occurrence, by condition_concept_id
14. Number of condition occurrence records, by condition_concept_id
15. Number of condition_occurrence records by condition_source_concept_id
16. Number of persons with at least one procedure occurrence, by procedure_concept_id
17. Number of procedure occurrence records, by procedure_concept_id
18. Number of procedure_occurrence records by procedure_source_concept_id
19. Number of persons with at least one drug exposure, by drug_concept_id
20. Number of drug exposure records, by drug_concept_id
21. Number of drug_exposure records by drug_source_concept_id
22. Number of persons with at least one observation occurrence, by observation_concept_id
23. Number of observation occurrence records, by observation_concept_id
24. Number of observation records by observation_source_concept_id
25. Number of persons with at least one drug era, by drug_concept_id
26. Number of drug era records, by drug_concept_id
27. Number of persons with at least one measurement occurrence, by measurement_concept_id
28. Number of measurement occurrence records, by measurement_concept_id
29. Number of measurement records with no value (numeric, string, or concept)
30. Distribution of numeric values, by measurement_concept_id and unit_concept_id
31. Number of measurement records by measurement_source_concept_id
32. Number of distinct patients that overlap between specific domains
33. Number of persons with at least one device exposure, by device_concept_id
34. Number of device exposure records, by device_concept_id
35. Number of device_exposure records by device_source_concept_id

APPENDIX 2

Full list of Database Diagnostics settings objects used in this study

This appendix will be furnished upon request.

SUPPLEMENTARY MATERIAL

The supplementary data resulting from this study will be furnished at the request of the reader.

DECLARATIONS

Author Contributions

CB, MAS, TP were involved in data collection. CB, MJS, PBR, MAS, PRR were involved in the study design, analysis, and interpretation of results. CB, MJS, PRR contributed to writing and all authors revised and approved the final draft.

Competing Interests

CB, MJS, PBR, are employees of Johnson & Johnson and hold stock and stock options. PRR works for a department that receives/received unconditional research grants from Amgen, Chiesi, Johnson and Johnson, UCB Biopharma, the European Medicines Agency and the Innovative Medicines Initiative.

Patient Consent

Not required

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

1. Fleurence RL, Curtis LH, Califf RM, *et al.* Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*. 2014;21:578–82. doi: 10.1136/amiajnl-2014-002747
2. Brown JS, Holmes JH, Shah K, *et al.* Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Medical Care*. 2010;48:S45–51. doi: 10.1097/MLR.0b013e3181dbdcdb
3. Ohmann C, Banzi R, Canham S, *et al.* Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open*. 2017;7:e018647. doi: 10.1136/bmjopen-2017-018647
4. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574–8.
5. Rijnbeek PR, Schuemie MJ, van der Lei J, *et al.* The European Health Data & Evidence Network (EHDEN): building a federated network to accelerate research. *European Journal of Epidemiology*. 2020;35:613–7. doi: 10.1007/s10654-020-00606-5
6. Voss EA, Shoaibi A, Yin Hui Lai L, *et al.* Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 databases: a multinational retrospective cohort study. *EClinicalMedicine*. 2023;58:101932. doi: 10.1016/j.eclinm.2023.101932
7. Li X, Ostropolets A, Makadia R, *et al.* Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. *BMJ*. 2021;373:n1435. doi: 10.1136/bmj.n1435
8. European Medicines Agency. Darwin EU. https://www.darwin-eu.org/ (accessed 26 September 2024)
9. Blacketer C, Defalco FJ, Ryan PB, *et al.* Increasing trust in real-world evidence through evaluation of observational data quality. *Journal of the American Medical Informatics Association : JAMIA*. Published Online First: 27 July 2021. doi: 10.1093/jamia/ocab132
10. Kahn MG, Callahan TJ, Barnard J, *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016;4:1244. doi: 10.13063/2327-9214.1244
11. Callahan TJ, Bauck AE, Bertoch D, *et al.* A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *EGEMS (Wash DC)*. 2017;5:8. doi: 10.5334/egems.223
12. Weiskopf NG, Bakken S, Hripcsak G, *et al.* A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC)*. ;5. doi: 10.5334/egems.218
13. Kahn MG, Brown JS, Chun AT, *et al.* Transparent reporting of data quality in distributed data networks. *EGEMS (Washington, DC)*. 2015;3:1052. doi: 10.13063/2327-9214.1052
14. Dixon BE, Wen C, French T, *et al.* Extending an open-source tool to measure data quality: case report on Observational Health Data Science and Informatics (OHDSI). *BMJ Health Care Inform*. 2020;27. doi: 10.1136/bmjhci-2019-100054
15. Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences of the United States of America*. 2016;113:7329–36. doi: 10.1073/pnas.1510502113
16. Voss EA, Makadia R, Matcho A, *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association*. 2015;22:553–64. doi: 10.1093/jamia/ocu023

17. Observational Health Data Sciences and Informatics. OMOP Common Data Model (CDM). https://ohdsi.github.io/CommonDataModel/
18. Reich C, Ostropolets A, Ryan P, *et al.* OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*. 2024;31:583–90. doi: 10.1093/jamia/ocad247
19. Rao G, Shoaibi A, Makadia R, *et al.* CohortDiagnostics: phenotype evaluation across a network of observational data sources using population-level characterization. *medRxiv*. Published Online First: 2023. doi: 10.1101/2023.06.28.23291982
20. Conover MM, Ryan PB, Chen Y, *et al.* Objective study validity diagnostics: a framework requiring pre-specified, empirical verification to increase trust in the reliability of real-world evidence. *Journal of the American Medical Informatics Association*. 2025;32:518–25. doi: 10.1093/jamia/ocad030
21. Schuemie MJ, Ryan PB, Pratt N, *et al.* Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study. *J Am Med Inform Assoc*. 2020;27:1268–77. doi: 10.1093/jamia/ocaa124
22. Suchard MA, Schuemie MJ, Krumholz HM, *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet*. 2019;394:1816–26. doi: 10.1016/S0140-6736(19)32317-7
23. Khera R, Aminorroaya A, Dhingra LS, *et al.* Comparative Effectiveness of Second-Line Antihyperglycemic Agents for Cardiovascular Outcomes: A Multinational, Federated Analysis of LEGEND-T2DM. *Journal of the American College of Cardiology*. 2024;84:904–17. doi: 10.1016/j.jacc.2024.05.069

General Discussion

Clair Blacketer

Statistical approaches and technology to support evidence generation will continue to change, but data remains the foundation upon which all methodologies, assumptions, and study designs ultimately rest. This thesis seeks to strengthen that foundation through investigation and improvement of data standardization, data quality, and fitness-for-purpose of observational health data. The following sections summarize key insights from this research and outline directions for future work to advance each of these ideas.

The term *data quality* is frequently raised in academic and regulatory discussions, yet it is more vague than it first appears [1–3]. Few would dispute its importance, but the field is yet to agree on a shared operational definition. Numerous frameworks have attempted to formalize the concept, including those developed by regulators to guide the use of real-world data in decision making [4–9]. These definitions often intersect with *data standardization* and *fitness-for-purpose,* the two other domains explored in this thesis. Standardization provides the structural and semantic foundation to evaluate data quality in a consistent way, while fitness-for-purpose reflects how well that data can answer a specific research question, or how quality data informs studies. Therefore, interpreting and advancing data quality requires understanding the relationship between how data are structured, evaluated, and eventually used to generate evidence. Through this work, several definitions of data quality have surfaced, each raising further questions and thoughts about how these dimensions can be harmonized to support reliable, reproducible research.

DATA STANDARDIZATION

Are all the data from the source represented in the CDM?

Concerns over data loss are common when a data source is first converted to a common model. These concerns are valid, especially when records are reorganized or source codes are split into multiple standard concepts, making one-to-one comparisons between native and standardized data difficult. Study replications are often conducted to reassure data owners and users that no meaningful information was lost, though such efforts can expend valuable time and resources [10].

In practice, the extract, transform, and load (ETL) process used to standardize data also includes many data quality remediations [11–13]. The goal of this process is not just reorganization of data but also the removal of inconsistencies and errors, making the standardized version of a database a clean analytic layer. While effectively taking the responsibility from the researcher, this habit also moves the standardized data even farther from the native data in comparison.

Consider a patient record with an implausible birth year of 1850. If left unaddressed, researchers would need to identify and exclude that record in every subsequent analysis. When removed during the ETL, the problem is solved in all downstream uses of the data. As the ETL process is iterative and cumulative, each cycle of improvement adds more logic to handle data quality issues systematically. While this approach is efficient, it can have unintended consequences. The standardized version of a database can gradually differ from the native, obscuring how much cleaning was done. Replication studies may then produce different results, not because the true effect has changed, but because the data quality remediation removed implausible or low-quality records that were previously included. This can erode confidence in the data when, in reality, what has changed is the removal of "junk". This practice is often criticized as *data loss,* a term with negative connotations that evokes skepticism and distrust [14]. But that interpretation misses the important point that the removal of implausible or inconsistent data can improve reliability if the process is well-documented and reproducible. Trust in data is lost not through acts of remediation, but as a product of the uncertainty that can arise through an opaque process of removal. Increasing transparency in this process would strengthen confidence both in the ETL process and the resulting standardized data.

Future Directions

Future research in this space should focus on updating the ETL process to separate the reorganization of data from vocabulary mapping as well as data quality assessment and remediation. One potential approach would be to treat these as three distinct phases of data standardization. The first could be an initial structural mapping phase in which all records from the source data are moved into the OMOP CDM tables without modification. This would allow for direct comparisons between the source and standardized datasets to show that patient counts, code frequencies, and other key metrics are equivalent between the two.

Once this structural mapping is complete, vocabulary mapping could then follow as the second phase. Since the source data are already in a standard structure, universal software for handling vocabulary mapping and subsequent domain movement could be developed. Transparent attrition reporting that documents where and why records were excluded or moved would alleviate concerns about data loss by demonstrating that the initial mapping preserved all available information prior to the application of standard concepts.

After the structural and vocabulary mapping phases are complete, quality assessment and remediation could follow as the third phase. Applying the Data Quality Dashboard (DQD) to the fully standardized but uncleaned data would expose the full picture of data quality issues within a database. Universal remediation logic could then be developed, shared, and reused across sites, reducing redundancy and promoting consistency in

how quality issues are addressed. This three-step approach could increase clarity, enable more reproducible workflows, and further lower the barriers for institutions interested in adopting the OMOP CDM. In addition, the three distinct phases of this suggested ETL update would produce multiple versions of the same data, supporting continued methods research into how data standardization potentially influences study results.

Certain characteristics of data cannot be easily standardized, even with greater transparency in ETL processes. One problem that comes up frequently is how to define observable time, or the time during which a patient's clinical events are expected to be reliably captured in a database. Chapter 2 showed that different definitions of the observation period across databases can impact incidence rate estimates, making it a methodological choice of great consequence. In spite of its importance, no single definition is universally acceptable; data structures and the underlying assumptions that accompany them differ greatly between and among databases. Recent work, such as that presented by Andy South and colleagues, has suggested that observation periods in electronic health record data should be defined using a "first record to extraction" approach as default [15]. However, if such an approach is widely adopted, study results should be stratified by clinical context to account for inherent differences in data capture. Future research could build on these ideas to develop patient-level prediction models to determine appropriate observation periods based on expected utilization patterns per person.

DATA QUALITY

Does my database represent the true state of the patient?

This question is possibly the most common definition of observational health data quality [4–6]. To generate reliable evidence our data should faithfully represent a patient's clinical reality, a target that is difficult to achieve and even harder to codify. Most data quality frameworks in use are specifically aimed at this idea, focusing on how well the data mirror the true state of the patient.

Aligning with the Kahn framework, many federated networks, including those using the OMOP CDM, are strongest at assessing conformance, moderately capable at measuring completeness, and least comprehensive in evaluating plausibility [2,16]. Conformance is the most straightforward, as it verifies whether data adhere to the structural and semantic expectations of the model. Completeness, however, is more difficult as a single database cannot be expected to contain every detail of a patient's full medical journey, and evaluating its quality through that lens is impractical. Instead, completeness should be assessed through an understanding of the clinical context represented. For example, a hospital-based electronic health record will naturally include long periods where patients

have no recorded activity, whereas an administrative claims database containing outpatient pharmacy data would be expected to show more consistent longitudinal coverage.

Existing initiatives provide helpful examples of how measurement of data completeness can be operationalized. The Patient Centered Outcomes Research Network (PCORNet) quality curation process applies predefined thresholds ensure the data comply with expected density measures. For example, encounters are assessed to ensure that the average number of procedures per encounter are at least 1.0. Anything less than this number could indicate a missing data stream or data mapping errors [2]. In the context of PCORNet, which is comprised of electronic health record (EHR) data from academic medical centers in the United States, this requirement aligns with expectations of patient care in those settings. Similarly, the FDA Sentinel Initiative's data quality completeness module checks that required fields like admission and discharge dates are consistently populated [16]. US administrative claims make up the majority of data within Sentinel, where these dates must be captured for reimbursement purposes.

Plausibility remains the most conceptually and operationally difficult dimension to evaluate. PCORNet and Sentinel again differ in their approach, with PCORNet focusing more on correcting height, weight, and blood pressure values and Sentinel concentrating on date errors to ensure that hospital admissions occur during a patient's enrollment in an insurance plan [2,16]. In OMOP CDM standardized databases, the Data Quality Dashboard (DQD) can identify values or relationships that appear implausible as described in Chapter 3 of this thesis. Across these approaches, there is little consensus on how to interpret or resolve such findings. More fundamentally, the field still lacks a clear understanding of how different dimensions of data quality, in particular plausibility, affect the evidence ultimately generated.

Setting context-appropriate expectations for both data completeness and plausibility across diverse data sources (rather than applying uniform benchmarks) continues to be an important process. Building on Patti Brennan's definitions of data quality as the degree to which data represent the true state of the patient, this thesis adds the important qualification that data should represent the true state of the patient *within the given clinical context* [17]**.** Quality is not an isolated attribute of a database; it is a natural consequence of purpose and provenance.

Does my data look like other databases with similar clinical contexts?

Data partners are often curious about this question. Progress in both standardization and quality assessment has been significant, but the deeper understanding of how one database compares to another (even when the databases represent similar clinical contexts) remains incomplete. Establishing a framework to define and structure these comparisons would help drive further network quality improvement by quickly and clearly pointing

out potential areas of incompleteness or disagreement (e.g. comparing specific rates of conditions, conditions, procedures, or drug exposures.), whether due to population variation, coding practices, or data capture processes.

Future Directions

While tools such as the DQD have made it possible to systematically identify potential data quality issues, there is less certainty about how these issues should be interpreted. For example, implausible temporal relationships like records after death can be detected, but their meaning and impact on evidence generation remains unclear. Future research could focus on developing methods to better understand these anomalies. For instance, we currently lack consistent approaches to determine whether implausible values reflect true data entry errors, rare but legitimate clinical events, or artifacts of data capture processes.

Equally important, there is limited consensus on how to remediate these inconsistencies in ways that maintain both data integrity and provenance. Using the previous example, should all records that occur after a patient's death be removed from the database, assuming they are all errors? What if they represent potentially expected events after death, like an autopsy? Is a better approach to remove all records that occur more than a certain number of days after death? How many days? More research is needed to determine the best approach to validate implausible data at the patient level, as well as empirical, transparent strategies to address true data quality issues that could be applied across databases. Such efforts may also include evaluating how data quality interventions influence downstream evidence generation. Building the empirical link between data quality and inference would not only strengthen methodological rigor but also provide a more nuanced understanding of what "high-quality data" means in the context of observational health research.

The next opportunity in data quality involves the development of learning systems, or dynamic frameworks capable of continuous data improvement through feedback loops between sites and a central coordinating center. Centralized networks like the National Covid Cohort Collborative (N3C) and PEDSNet have demonstrated the value and application of iterative quality improvement, but extending this approach to federated networks introduces new challenges and opportunities [18–20]. A federated learning system for data quality should enable sites to share knowledge without sharing patient-level data. By learning from experiences of diverse databases, this system could work towards refining quality expectations that consider the clinical context of data rather than being uniform across all sources. In doing so, federated learning could support a shift from a one-size-fits-all approach towards *context-dependent* data quality measures.

Similarly, new methods are needed that will allow for meaningful, comprehensible comparisons between data sources. Such comparisons could reveal differences in data capture, coding practices, and population characteristics, supporting both quality improvement and better-informed study planning. For example, benchmark analyses might identify gaps or strengths by looking at variations in condition prevalence, procedure rates, or drug utilization patterns across databases representing similar care settings. Over time, this type of benchmarking could help establish expectations for what typical data in a given clinical context looks like.

FITNESS FOR PURPOSE

Is my data fit to generate evidence in support of a specific scientific question?

Beyond evaluating the presence of necessary elements for a given study, assessing fitness-for-purpose requires understanding whether a database can support the *type* of study being conducted. Not all studies impose the same demands on data. Most databases can support characterization analyses, which describe patterns of care, utilization, or outcomes within a population. However, estimation and prediction studies introduce additional expectations and assumptions, one of the most critical being temporality. A database must accurately reflect when clinical events occurred in relation to one another to support causal inference or predictive modeling. Without reliable temporal information, even a database rich in content may not be fit for these types of analyses.

Future Directions

Central to this effort and others described above is the need to categorize databases by clinical context to ensure that such comparisons are valid and interpretable. Chapter 5 suggested the dimensions of person inclusion, care level, and data capture method as possible organizing structures. Other frameworks have been proposed in the literature, but advancing this work will require greater consensus on which approach to use. Establishing a consistent way to define and classify clinical context is the necessary first step toward evaluating how well a database reflects real-world care and how it compares to others with similar contexts [21,22].

In addition, measurable database characteristics could help shape and inform the types of studies that databases can reasonably support. Characteristics such as data heterogeneity, granularity, temporality, timeliness, and domain coverage can improve Database Diagnostics by going beyond determining if a database has the required building blocks for a study. For instance, a very granular database with detailed, temporal records of clinical encounters might be best for estimation or prediction studies, while a less granular data source could be more appropriate for descriptive characterization. Additional

research is needed to help understand precisely how variation in these characteristics across databases play out in study results and how they can be used to better assess database fitness-for-purpose.

These efforts could move the field toward a more systematic understanding of fitness-for-purpose, allowing researchers to anticipate which databases can meaningfully contribute not only to a given study, but to a given study *design* as well. In doing so, they would advance both the reliability and the interpretability of federated research, ensuring that each data source is used where it is most capable of informing evidence generation.

CONCLUSION

In her "Notes on Hospitals", Florence Nightingale articulated many of the same challenges and ideas explored in this thesis [23].

"The data at my disposal are too limited to allow of the comparative effect of age and sex on mortality being discovered; but, incomplete as they are, they all go to prove the necessity for adopting a uniform nomenclature of complications, and a uniform method of tabulation" [23].

She bemoaned the difficulty of comparing patient outcomes, particularly surgical complications, because of the inconsistent nature of the data available. The same condition might be recorded under several names, and outcomes were often defined differently across hospitals. To address this, Nightingale proposed a uniform nomenclature for surgical procedures and outcomes, along with standard reporting forms to ensure comparability. Recognizing that data standardization and quality were prerequisites for learning from clinical experiences, she made an "urgent appeal for adopting this or some *uniform* system of publishing the statistical records of hospitals." Her next sentence continues: "There is growing conviction that in all hospitals, even in those which are best conducted, there is a great and unnecessary waste of life…", linking her call for uniform data directly to the improvement of patient care [23].

Nightingale's final entry on the use of data to impact patient lives remains remarkably prescient. She posited that, with reliable hospital records, "we should be able to decide many other questions…of the relative value of particular operations and modes of treatment than we have any means of ascertaining at present." Her vision was that observational health data, standardized and used appropriately, would "save life and suffering, and improve the treatment and management of the sick and maimed poor." [23]

More than a century and a half later, the medical informatics community continues to pursue this same goal. The statistical methods and technologies have changed over time, but generating trustworthy evidence from imperfect data remains a core need. The work described in this thesis builds upon the same foundation Nightingale suggested, one based in standardization, quality, and fitness-for-purpose. By improving how we structure, assess, and evaluate observational health data, we take steps toward realizing her vision on a global scale. Yet, as Nightingale reminded us, the aim of these efforts is not just technical precision, but human benefit. It may begin with the data, but it ends with the patient.

REFERENCES

1. Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Ann Intern Med*. 2020;172:S79–84. doi: 10.7326/M19-0873
2. Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMS (Wash DC)*. ;6. doi: 10.5334/egems.199
3. Alla F, Rosilio M, Funck-Brentano C, *et al.* How can the quality of medical data in pharmacovigilance, pharmacoepidemiology and clinical studies be guaranteed? *Therapie*. 2013;68:209–23. doi: 10.2515/therapie/2013040
4. Bian J, Lyu T, Loiacono A, *et al.* Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc*. 2020;27:1999–2010. doi: 10.1093/jamia/ocaa245
5. Johnson SG, Speedie S, Simon G, *et al.* A Data Quality Ontology for the Secondary Use of EHR Data. *AMIA Annu Symp Proc*. 2015;2015:1937–46.
6. Heads of Medicines Agency, European Medicines Agency. Data Quality Framework for EU Medicines Regulation. 2023.
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20:144–51. doi: 10.1136/amiajnl-2011-000681
8. Burns L, Roux NL, Kalesnik-Orszulak R, *et al.* Real-World Evidence for Regulatory Decision-Making: Guidance From Around the World. *Clin Ther*. 2022;44:420–37. doi: 10.1016/j.clinthera.2022.01.012
9. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018.
10. Candore G, Hedenmalm K, Slattery J, *et al.* Can We Rely on Results From IQVIA Medical Research Data UK Converted to the Observational Medical Outcome Partnership Common Data Model? *Clin Pharmacol Ther*. 2020;107:915–25. doi: 10.1002/cpt.1785
11. Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *EGEMS (Wash DC)*. 2014;2:1110. doi: 10.13063/2327-9214.1110
12. Biedermann P, Ong R, Davydov A, *et al.* Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med Res Methodol*. 2021;21:238. doi: 10.1186/s12874-021-01434-3
13. Voss EA, Makadia R, Matcho A, *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. 2015;22:553–64. doi: 10.1093/jamia/ocu023
14. Painter JL, Ramcharran D, Bate A. Perspective review: Will generative AI make common data models obsolete in future analyses of distributed data networks? *Ther Adv Drug Saf*. 2025;16:20420986251332743. doi: 10.1177/20420986251332743
15. Andy South, Hiba Junaid, Jennifer Lane, *et al.* Moving towards standard observation period settings for OMOP-based EHR data: results from the HERON-UK network. *OHDSI Europe Symposium 2025*.
16. U.S. Food and Drug Administration. An Assessment of the Sentinel System (2022 to 2024). 2025.
17. Brennan PF, Stead WW. Assessing Data Quality. *J Am Med Inform Assoc*. 2000;7:106–7.

18. Pfaff ER, Girvin AT, Gabriel DL, *et al.* Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. *J Am Med Inform Assoc*. 2022;29:609–18. doi: 10.1093/jamia/ocab217
19. Khare R, Utidjian LH, Razzaghi H, *et al.* Design and Refinement of a Data Quality Assessment Workflow for a Large Pediatric Research Network. *EGEMS (Wash DC)*. ;7. doi: 10.5334/egems.294
20. Kahn MG, Brown JS, Chun AT, *et al.* Transparent reporting of data quality in distributed data networks. *EGEMS (Washington, DC)*. 2015;3:1052. doi: 10.13063/2327-9214.1052
21. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*. 2022;22:287. doi: 10.1186/s12874-022-01768-6
22. Gini R, Pajouheshnia R, Gardarsdottir H, *et al.* Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review. *Pharmacoepidemiol Drug Saf*. 2024;33:e5787. doi: 10.1002/pds.5787
23. Nightingale F. *Notes on Hospitals*. 3rd ed. London: Longman, Green, Longman, Roberts, and Green 1863.

SUMMARY

Well known for her heroic efforts in the Crimean war and innovations in nursing practice, Florence Nightingale was also a gifted statistician. She believed that data collected from London hospitals could and should be used to generate evidence in pursuit of improved patient care. In her work "Notes on Hospitals" she expands on this idea, describing how the future of medical informatics hinges on the need for standardized terminology and data structure. Today, the scope of data has expanded from hospitals in one urban city to databases around the world representing millions of patients, but the need for high-quality data to support research has not changed. This thesis focuses on improving the foundational science of data standardization, data quality, and fitness-for-purpose in federated networks to increase trust in evidence and ultimately improve patient lives.

Data Standardization

Chapter 1 examines the practicalities of standardizing diverse data sources to the OMOP Common Data Model (CDM) through the experience of building the European Health Data and Evidence Network (EHDEN). Anecdotally, we knew the challenges and pitfalls of the extract, transform, and load (ETL) process, but this study allowed us to observe and support twenty-five data partners from 11 countries as they standardized their data. The largest factor that contributed to a data partners' success was team composition. Such groups should be balanced with individuals with deep knowledge of the source data, individuals with expertise in the OMOP CDM and vocabularies, and individuals with the technical skills to implement the data transformation. The most consistent hurdle for teams was vocabulary mapping, where source codes that could not be automatically assigned to standard concepts required manual review. Large language models are now being leveraged to reduce this burden, and the introduction of community-driven mappings by the OHDSI vocabulary team marks an important step toward a more adaptive and collaborative future for vocabulary standardization. ETL teams also wrestled with issues of granularity, equivalency, and the practical question of how much effort to invest in resolving rare or ambiguous codes, though these challenges can often be mitigated through use case–driven mapping strategies that focus effort where it has the greatest analytical relevance. Finally, this work highlighted the need for clear data governance agreements prior to the initiation of data standardization. Membership in a federated network presupposes the participation of the data partner in network research. Governance requirements to allow such participation should be clear and unambiguous. Without clarity, barriers can emerge later during analysis or publication.

Chapter 2 showcases an opportunity to implement standard methodologies to address assumptions about the underlying population of a database that were previously handled at the time of analysis. 80 definitions of observable time across 11 databases were used to generate incidence rates of five different outcomes of interest. The analysis

showed that even relatively minor differences in defining when observation begins and ends for a patient can materially alter incidence rate estimates. Importantly, standardization offers a mechanism to implement consistent approaches to observable time across diverse databases while still accommodating variation in data capture methods.

Data Quality

Standardization not only enables methodological innovation and federated research but also creates the conditions for systematic quality assessment. While scientists can design and test studies without accessing patient-level data, data owners must still execute analyses and return results to study leads. For this process to be trustworthy, there must be a way to evaluate whether each database meets minimum quality standards without compromising patient privacy. **Chapter 3** describes the development of the Data Quality Dashboard (DQD) to meet this need. The standardized scaffold of the OMOP CDM provides a foundation for implementing quality checks at scale, allowing data partners to generate transparent, reproducible assessments of their databases. Results can then be communicated to coordinating bodies without exposing patient-level information. Requiring databases mapped to the CDM to run the DQD represents an important first step toward consistently reporting data quality, thereby strengthening confidence in the data used to support observational studies in federated networks.

Chapter 4 then utilizes the data standardization exercises taking place among data partners in EHDEN to measure how well the DQD improves the quality of databases in federated networks. The tool was applied to 25 databases mapped to the OMOP CDM spanning diverse data origins, data capture methods, and underlying populations. Requiring partners to complete the DQD ensured that each database met baseline quality expectations before participating in federated studies. In particular, the tool performed well in assessing conformance to CDM structural requirements, providing assurance that standardized study packages could run without error. While study-specific checks remain essential, the DQD established a consistent foundation of conformance, completeness, and plausibility across the network. This application demonstrated the value of the DQD not only for individual data partners but also for federated networks, ensuring the data are ready when the needs arise.

Fitness for Purpose

Chapter 5 gives an overview of the more than 200 data sources from 29 countries participating in the EHDEN network. The breadth of healthcare systems, data capture methods, and populations represented across these sources underscores the importance of drawing on multiple datasets to capture the full range of patient experiences in Europe. At the same time, this diversity highlights the persistent challenge that no single database can be expected to support every (or even most) research questions posed to a network. Determining whether a database is fit-for-purpose therefore is an essential step before

initiating studies. Relying on site-level feasibility assessments based on patient-level data is neither sustainable nor scalable, given the time and resource demands. As federated networks expand in both size and visibility, including within regulatory initiatives such as DARWIN EU®, new approaches are needed to assess database fitness-for-purpose in a way that is both efficient and privacy-preserving.

Chapter 6 offers a solution to this challenge by introducing and validating a method for performing early fit-for-purpose assessments in federated networks using only precomputed summary statistics. By eliminating the need to access patient-level data during study planning, this approach substantially reduces the time, resources, and governance burdens typically required at this stage. The method, known as Database Diagnostics, goes beyond a simple yes-or-no feasibility check by providing diagnostic insights into why a particular study may not be viable in a given database. These insights can guide both data curation efforts and refinements in study design, improving the efficiency and interpretability of network research. In doing so, Database Diagnostics fills an important methodological and operational gap, offering a scalable and privacy-preserving way to evaluate database readiness and ultimately strengthen the generation of real-world evidence across federated networks.

❄ ❄ ❄

This thesis drew parallels to the historical work of Florence Nightingale to illustrate that the needs for high-quality standardized data are not new to the field of medical informatics. However, with the unprecedented amount of data and technologies available, we have both the opportunity and obligation to improve the science of data standardization, quality, and fitness-for-purpose to ensure the reliability of evidence. This thesis argues that these elements make up the foundation of federated observational health research. It is through their interrogation, application, and transparency that we as researchers can trust the science built on these foundations to provide actionable insights that improve patient care. It may start with the data, but it ends with the patient.

SAMENVATTING

Florence Nightingale staat bekend om haar heroïsche inspanningen in de Krimoorlog en innovaties in de verpleegkundige praktijk, maar was ook een begaafd statisticus. Ze was van mening dat de gegevens die werden verzameld in Londense ziekenhuizen konden en moesten worden gebruikt om bewijs te genereren dat bijdraagt aan verbeterde patiëntenzorg. In haar werk "Notes on Hospitals" gaat ze dieper in op dit idee en beschrijft ze hoe de toekomst van de medische informatica afhangt van de behoefte aan gestandaardiseerde terminologie en datastructuur. Tegenwoordig is de reikwijdte van de gegevens uitgebreid van ziekenhuizen in één stad naar databases over de hele wereld die miljoenen patiënten vertegenwoordigen, maar de behoefte aan hoogwaardige kwaliteit van gegevens ter ondersteuning van onderzoek is niet veranderd. Dit proefschrift richt zich op het verbeteren van de fundamentele wetenschap van standaardisatie van gegevens, gegevenskwaliteit en doelgeschiktheid in gefedereerde netwerken om het vertrouwen in bewijs te vergroten en uiteindelijk het leven van de patiënt te verbeteren.

Standaardisatie van gegevens

Hoofdstuk 1 onderzoekt de praktische aspecten van het standaardiseren van diverse gegevensbronnen naar het OMOP Common Data Model (CDM) aan de hand van de ervaring van het bouwen van het European Health Data and Evidence Network (EHDEN). Anekdotisch kenden we de uitdagingen en valkuilen van het extract, transform en load (ETL) proces, maar dankzij dit onderzoek konden we vijfentwintig datapartners uit 11 landen observeren en ondersteunen bij het standaardiseren van hun gegevens. De grootste factor die bijdroeg aan het succes van een datapartner was de samenstelling van het team. Bij het samenstellen van het team moet een balans gevonden worden tussen verschillende expertises: grondige kennis van de brongegevens, expertise in het OMOP CDM en vocabulaires en technische vaardigheden om de gegevenstransformatie te implementeren. De meest voorkomende hindernis voor teams was het mappen van de vocabulaires, waarbij broncodes die niet automatisch konden worden toegewezen aan standaardconcepten handmatig moesten worden beoordeeld. Grote taalmodellen worden nu ingezet om deze last te verminderen, en de introductie van community-driven mappings door het OHDSI vocabulaire team is een belangrijke stap naar een meer adaptieve en collaboratieve toekomst voor standaardisatie van vocabulaires. ETL-teams worstelden ook met kwesties als granulariteit, gelijkwaardigheid en de praktische vraag hoeveel moeite te investeren in het oplossen van zeldzame of dubbelzinnige codes, hoewel deze uitdagingen vaak kunnen worden beperkt door middel van use case-gedreven mappingstrategieën waarbij de inspanning gericht is op de grootste analytische relevantie. Ten slotte benadrukte dit werk de noodzaak van duidelijke overeenkomsten rondom datagovernance voorafgaand aan de start van gegevensstandaardisatie. Deelname aan een gefedereerd netwerk veronderstelt immers de deelname van de datapartner aan netwerkonderzoek. De governancevereisten om een dergelijke deelname mogelijk te

maken moeten duidelijk en ondubbelzinnig zijn. Zonder duidelijkheid kunnen er later tijdens analyse of publicatie barrières aan het licht komen .

Hoofdstuk 2 laat zien dat het mogelijk is standaardmethodologieën te implementeren om aannames over de onderliggende populatie van een database te beoordelen die voorheen aangepakt werden op het moment van analyse. Uit 11 databases werden 80 definities van observeerbare tijd gebruikt om incidentiecijfers van vijf verschillende uitkomsten van interesse te genereren. De analyse toonde aan dat zelfs relatief kleine verschillen in het definiëren wanneer observatie begint en eindigt voor een patiënt, de schattingen van de incidentie aanzienlijk kunnen veranderen. Belangrijk is dat standaardisatie een mechanisme biedt om consistente benaderingen van observeerbare tijd te implementeren in verschillende databases, terwijl er nog steeds ruimte blijft voor variatie in methoden voor het vastleggen van gegevens .

Kwaliteit van gegevens

Standaardisatie maakt niet alleen methodologische innovatie en gefedereerd onderzoek mogelijk, maar schept ook de voorwaarden voor systematische kwaliteitsbeoordeling. Terwijl wetenschappers studies kunnen ontwerpen en testen zonder toegang te krijgen tot gegevens op patiëntniveau, moeten gegevenseigenaren nog steeds analyses uitvoeren en resultaten terugsturen naar onderzoeksleiders. Om dit proces betrouwbaar te maken, moet er een manier zijn om te evalueren of elke database voldoet aan minimale kwaliteitsstandaarden zonder de privacy van de patiënt in gevaar te brengen. **Hoofdstuk 3** beschrijft de ontwikkeling van het Data Quality Dashboard (DQD) om aan deze behoefte te voldoen. Het gestandaardiseerde raamwerk van het OMOP CDM biedt een basis voor het uitvoeren van kwaliteitscontroles op grote schaal, waardoor datapartners transparante, reproduceerbare beoordelingen van hun databases kunnen genereren. De resultaten kunnen vervolgens worden gedeeld met de coördinerende instanties zonder dat informatie op patiëntniveau wordt onthuld. Het vereisen dat databases die het CDM gebruiken de DQD uitvoeren is een belangrijke eerste stap richting consistente rapportage van gegevenskwaliteit, waardoor het vertrouwen in de gegevens die worden gebruikt in observationele studies in gefedereerde netwerken wordt versterkt.

Hoofdstuk 4 maakt vervolgens gebruik van de standaardisatie van gegevens die plaatsvindt tussen datapartners in EHDEN om te meten hoe de DQD de kwaliteit van databases in gefedereerde netwerken verbetert. De tool is toegepast op 25 databases die zijn geconverteerd naar het OMOP CDM en die verschillende gegevensbronnen, methoden voor gegevensvastlegging en onderliggende populaties omvatten. Door van datapartners te eisen om de DQD te voltooien, werd ervoor gezorgd dat elke database voldeed aan de basiskwaliteitsverwachtingen voordat ze deelnamen aan gefedereerde studies. De tool presteerde met name goed bij het beoordelen van de conformiteit aan de structurele eisen van het CDM en bood ook de zekerheid dat gestandaardiseerde

studiepakketten foutloos konden worden uitgevoerd. Hoewel studiespecifieke controles essentieel blijven, heeft de DQD een consistente basis gelegd van conformiteit, volledigheid en plausibiliteit in het hele netwerk. Deze applicatie toonde de waarde van de DQD niet alleen voor individuele datapartners, maar ook voor gefedereerde netwerken, waardoor de gegevens klaar zijn als ze nodig zijn.

Geschiktheid voor het doel

Hoofdstuk 5 geeft een overzicht van de meer dan 200 gegevensbronnen uit 29 landen die deelnemen aan het EHDEN netwerk. De verscheidenheid van gezondheidszorgsystemen, methoden voor gegevensverzameling en populaties die in deze bronnen vertegenwoordigd zijn, benadrukt het belang van het gebruik van meerdere datasets om inzicht te geven in het volledige scala aan patiëntervaringen in Europa. Tegelijkertijd benadrukt deze diversiteit de uitdaging dat van geen enkele database kan worden verwacht dat deze elke (of zelfs de meeste) onderzoeksvragen ondersteunt die aan een netwerk worden gesteld. Bepalen of een database geschikt is voor het beoogde doel is daarom een essentiële stap voordat studies worden gestart. Vertrouwen op haalbaarheidsbeoordelingen per locatie op basis van gegevens op patiëntniveau is niet duurzaam of schaalbaar, gezien de benodigde tijd en middelen. Aangezien gefedereerde netwerken zowel qua omvang als qua zichtbaarheid toenemen, ook binnen regelgevende initiatieven zoals DARWIN EU®, zijn nieuwe benaderingen nodig om de geschiktheid van databases voor het beoogde doel te beoordelen op een manier die zowel efficiënt is als de privacy beschermt.

Hoofdstuk 6 biedt een oplossing voor deze uitdaging door een methode te introduceren en te valideren voor het uitvoeren van vroege doelgeschiktheidsbeoordelingen in gefedereerde netwerken gebruikmakend van alleen vooraf berekende samenvattende statistieken. Door het elimineren van de noodzaak om toegang te krijgen tot gegevens op patiëntniveau tijdens de studieplanning, vermindert deze aanpak aanzienlijk de tijd, middelen en governancelasten die in deze fase doorgaans nodig zijn. De methode, bekend als Database Diagnostics, gaat verder dan een eenvoudige ja-of-nee haalbaarheidscontrole door diagnostische inzichten te bieden in waarom een bepaald onderzoek al dan niet mogelijk is in een bepaalde database. Deze inzichten kunnen als leidraad dienen voor zowel inspanningen op het gebied van datacuratie als verfijningen in de onderzoeksopzet, waardoor de efficiëntie en interpreteerbaarheid van netwerkonderzoek wordt verbeterd. Database Diagnostics vervult hierdoor een belangrijke methodologische en operationele leemte op door een schaalbare en privacybeschermende manier om de gereedheid van de database te evalueren en uiteindelijk het genereren van real-world bewijs in gefedereerde netwerken te versterken.

❄ ❄ ❄

Dit proefschrift trok parallellen met het historische werk van Florence Nightingale om te illustreren dat de behoefte aan gestandaardiseerde gegevens van hoge kwaliteit niet nieuw is op het gebied van medische informatica. Met de ongekende hoeveelheid beschikbare gegevens en technologieën hebben we echter zowel de mogelijkheid als de verplichting om de wetenschap van gegevensstandaardisatie, kwaliteit en geschiktheid voor het doel te verbeteren om zo de betrouwbaarheid van bewijs te waarborgen. Dit proefschrift stelt dat deze elementen het fundament vormen van gedegen gefedereerd observationeel gezondheidsonderzoek. Het is door kritische toetsing, toepassing en transparantie hiervan dat we als onderzoekers kunnen vertrouwen op de wetenschap die op deze fundamenten is gebouwd om bruikbare inzichten te bieden die de patiëntenzorg verbeteren. Het begint wellicht met de gegevens, maar het eindigt bij de patiënt.

ACKNOWLEDGEMENTS

This work is the result of the support, guidance, and generosity of many people, and I am deeply grateful to all who have been part of this journey.

To my parents, **Meg and Rick Lemaster**, thank you for always believing in me. Mom, I often think back to our early days together, and if you had asked me then, I would not have imagined I could come this far. You sacrificed so much to give me the start I had, and I carry that with deep gratitude. Dad, you have always been unwaveringly kind and openly proud of me. Your support means more than I can express, especially because you chose to take me on as your own when you didn't have to. That choice has shaped my life in ways I will never fully be able to repay.

To my grandmothers, **Carol Ferrara and Cheryl Farrer**, you have been to every performance, recital, game, and graduation I've ever had. Your support is a constant and I know I can always count on you for words of encouragement. You have always been a steady presence in my life, and I carry that with me in everything I do. **Grandma**, I hope I can continue to be a source of pride for you as your faith and kindness are a constant for me. **Nana**, your curiosity and enthusiasm for learning inspire me to keep going and are a reminder to stay open, engaged, and curious.

To my extended family, **Aunt Vicki, Uncle Eric, Uncle Jack, Erica, Tyler**, thank you for always asking about my PhD, even when the answer rarely changed. Your continued interest and encouragement never went unnoticed. To my mother-in-law, **Sue Blacketer**, thank you for your kindness and for the steady positivity you have brought throughout this journey. Your support has meant a great deal to me.

To my Erasmus colleagues, thank you for always making time for me and for answering my many questions along the way. **Tineke De Ben** and **Ilse Schuemie**, your help navigating the ins and outs of the university was invaluable and made what could have been a complicated experience feel manageable. **Renske Los**, **Ross Williams**, **Victor Pera**, and **Tom Seinen**, thank you for always taking the time to chat about life, science, and the PhD journey. Those conversations made me feel like a real part of the team. You made what could have felt like a distant experience feel like home. **Aniek Markus**, you have been such a great friend throughout this time, helping me navigate the practical challenges and everything in between. I am truly grateful to have been part of such a supportive and thoughtful community.

To my Johnson & Johnson colleagues **Evanette Burrows, Jill Hardin, Mitch Conover, Rupa Makadia, Anthony Sena, Azza Shoaibi, Anna Ostropolets, Joel Swerdel, Dima Dymshyts,** and **Jenna Reps,** thank you for the constant encouragement and support

throughout this journey. You took the time to hear my ideas, read my papers, offer thoughtful feedback, and share suggestions that pushed this work further than I could have on my own. I am especially grateful for the positivity you brought to each step, it helped steady me through a long and often challenging experience. Your willingness to engage with this work made it better in ways that are reflected throughout this thesis. To **Frank DeFalco,** consider this your second honorary PhD. Thank you for your constant support, for helping bring ideas to life through visuals, and for the kindness you have always shown me, which pushed me to do better.

I am especially grateful to **Erica Voss**, who showed me this was possible by doing it herself. Watching you go through this gave me the confidence to believe I could do it too, and your example stayed with me on the days when I wasn't so sure. You faced it with strength, even through illness, and came out the other side stronger.

To my collaborators in the OHDSI community **Marc Suchard, Melanie Philofsky, Roger Carlson, Christian Reich, Asieh Golozar, Paul Nagy, Ben Hamlin, Ben Martin, Scott Duvall, Mui Van Zandt, and many others** thank you for your openness, your willingness to share ideas, and your commitment to building something bigger than any one of us. This work would not exist without the collective effort of this group. I am especially grateful for how thoughtful and supportive this community has been throughout this journey. It is rare to find a group that is not only advancing science, but doing so with such generosity, care, and shared purpose. I would like to thank **Maxim Moinat** and **Katy Sadowski,** who I have been fortunate to work closely with on data quality over the past few years. Your partnership, insight, and support have meant a great deal to me. This community has shaped not only this work, but how I think about collaboration and science more broadly.

To **Patrick Ryan,** thank you for always pushing me to be better and to think more critically. Your guidance has challenged me in the best ways and helped shape how I approach both my work and my thinking. You have a way of asking the right questions at the right time, and that has pushed me to go deeper and be more thoughtful in what I do. I am especially grateful that you encouraged me to pursue this opportunity at Erasmus. Your belief that I could do this gave me the confidence to take that step and see it through. I will carry what I have learned from you forward in the work that comes next as your influence is reflected in how I approach problems today.

To **Martijn Schuemie,** thank you for so graciously stepping in as my co-promotor partway through my PhD journey. Our regular meetings were a constant source of clarity and direction, and I am so grateful for the patience you showed as you listened, challenged, and guided me through the scientific method. Your feedback was always thoughtful and constructive, and you were incredibly generous with your time. I learned a great deal

from collaborating with you, and this work is stronger because of your guidance. You have helped shape how I think as a scientist, and I will carry that with me moving forward.

To **Peter Rijnbeek,** thank you for giving me the opportunity and space to grow into this work. Your kindness and steady support created an environment where I could grow and do my best. You have a way of encouraging without pressure, and of making the work feel meaningful and worthwhile. I always felt that you valued not just the outcome, but the thinking behind it, which pushed me to be more thoughtful and intentional in my work. I am grateful for the way you guided me through this journey, always helping me move forward while making me feel that what I was doing mattered. This experience has shaped not only my work, but how I approach research more broadly.

To my son, **James**, I am not sure there has been a time in your life when I have not been working on this PhD. I think fondly of the many evenings, weekends, and martial arts classes I have spent writing papers as you run, jump, kick, and fall around me. I am so grateful that I have been able to be present in your life and watch you grow while I pursue this goal. It almost feels as if we have grown together, you into an incredibly kind, sweet, happy kid, and me into a scientist in my own right. I cannot wait to see who you continue to become, and I hope you will go farther than I ever could.

To the little one who has been a quiet but constant companion during this final stretch. You are a reminder of what comes next, and I am so excited to meet you.

And to my husband, **Joey**, I have often said over the years that I wished I had a teacher like you. I am so proud of the work you do for your students, how you help them bring out their best selves, how you celebrate their successes as their own, and how you take their struggles to heart. Through this process, I feel like I have gotten a glimpse of what it is like to be your student. You have watched me write (or not write) and agonize over timelines and scientific choices. Not once did your faith in me waver though I questioned my own sanity multiple times. I still hear you in my head telling me to "WRITE YOUR PAPER" and, as silly it may sound, it was exactly what I needed. I don't think I have ever wanted to make someone as proud as I hope I make you. You see me more clearly than anyone else, and I truly do not think I could have done this without you.

CURRICULUM VITAE

Employment history

Johnson & Johnson Family of Companies
May 2015 - Present
City: Raritan, New Jersey
Country: United States of America

Anthem, Inc.
March 2014 – May 2015
City: Virginia Beach, Virginia
Country: United States of America

Sentara Health
May 2012 – March 2014
City: Virginia Beach, Virginia
Country: United States of America

Strelitz Diabetes Center
December 2010 – May 2012
City: Norfolk, Virginia
Country: United States of America

Education and training

Erasmus University Medical Center
PhD Health Data Science, Department of Medical Informatics
September 2019- Present

Eastern Virginia Medical School
Master of Public Health (concentration in Epidemiology)
September 2010 – August 2012

James Madison University
Bachelor of Science, Biology
September 2006 – May 2010

PHD PORTFOLIO

Description	Organizer	EC
Required		
Data Quality Presentation (2019)	Friends of the National Library of Medicine	1.00
OHDSI Global Symposium (2019)	OHDSI	2.00
OHDSI Community Calls	OHDSI	2.00
CBER BEST Data Quality Presentation	US FDA	1.00
Bio IT Conference Presentation	Bio IT	1.00
Digital Quality Summit (2021)	NCQA	1.00
OHDSI Global Symposium (2021)	OHDSI	2.00
RecoDID Presentation (2022)	RecoDID	1.00
OHDSI Europe Symposium (2022)	OHDSI	2.00
Digital Quality Summit (2022)	NCQA	1.00
OHDSI Global Symposium (2022)	OHDSI	2.00
Basel Epidemiology Research Seminar (2023)	BES	1.00
Scientific Integrity (2023)	Erasmus MC Graduate School	0.30
US National Academy of Sciences and Medicine Presentation	NASEM	1.00
OHDSI Global Symposium (2023)	OHDSI	2.00
Johns Hopkins University Data Quality Lecture	JHU	1.00
OHDSI Global Symposium (2025)	OHDSI	2.00
Total EC		30.30

LIST OF PUBLICATIONS

1. **Blacketer, C**., DeFalco, F. J., Conover, M. M., Ryan, P. B., Schuemie, M. J., & Rijnbeek, P. R. (2025). Evaluation of the impact of defining observable time in real-world data on outcome incidence. *Journal of the American Medical Informatics Association : JAMIA*, *32*(9), 1434–1444. https://doi.org/10.1093/jamia/ocaf119
2. **Blacketer, C**., Defalco, F. J., Ryan, P. B., & Rijnbeek, P. R. (2021). Increasing trust in real-world evidence through evaluation of observational data quality. *Journal of the American Medical Informatics Association : JAMIA*, *28*(10), 2251–2257. https://doi.org/10.1093/jamia/ocab132
3. **Blacketer, C**., Schuemie, M. J., Moinat, M., Voss, E. A., Camprubi, M., Rijnbeek, P. R., & Ryan, P. B. (2025). Advancing Real-World Evidence Through a Federated Health Data Network (EHDEN): Descriptive Study. *Journal of Medical Internet Research*, *27*, e74119. https://doi.org/10.2196/74119
4. **Blacketer, C**., Voss, E. A., DeFalco, F., Hughes, N., Schuemie, M. J., Moinat, M., & Rijnbeek, P. R. (2021). Using the Data Quality Dashboard to Improve the EHDEN Network. *Applied Sciences*, *11*(24), 11920. https://doi.org/10.3390/app112411920
5. **Blacketer, C**., Reps, J. M., Wang, L., Ryan, P. B., & Yuan, Z. (2022). Putting external validation performance of major bleeding risk models into context. *Frontiers in Drug Safety and Regulation*, *2*. https://doi.org/10.3389/fdsfr.2022.1034677
6. Bloom, L., Boyle, K. E., Myers, A. E., **Blacketer, C**., & Weinstein, R. (2019). Frequency of nonaspirin NSAID-relevant coexisting medical conditions in the primary-care setting: A retrospective database review. *Therapeutics and Clinical Risk Management*, *15*, 579–588. https://doi.org/10.2147/TCRM.S189833
7. Burn, E., Sena, A. G., Prats-Uribe, A., Spotnitz, M., DuVall, S., Lynch, K. E., Matheny, M. E., Nyberg, F., Ahmed, W.-U.-R., Alser, O., Alghoul, H., Alshammari, T., Zhang, L., Casajust, P., Areia, C., Shah, K., Reich, C., **Blacketer, C**., Andryc, A., … Duarte-Salles, T. (2021). Use of dialysis, tracheostomy, and extracorporeal membrane oxygenation among 842,928 patients hospitalized with COVID-19 in the United States. *medRxiv : The Preprint Server for Health Sciences*, 2020.11.25.20229088. https://doi.org/10.1101/2020.11.25.20229088
8. Cai, C. X., Nishimura, A., Bowring, M. G., Westlund, E., Tran, D., Ng, J. H., Nagy, P., Cook, M., McLeggon, J.-A., DuVall, S. L., Matheny, M. E., Golozar, A., Ostropolets, A., Minty, E., Desai, P., Bu, F., Toy, B., Hribar, M., Falconer, T., …**Blacketer, C**. … Ryan, P. B. (2024). Similar Risk of Kidney Failure among Patients with Blinding Diseases Who Receive Ranibizumab, Aflibercept, and Bevacizumab: An Observational Health Data Sciences and Informatics Network Study. *Ophthalmology. Retina*, *8*(8), 733–743. https://doi.org/10.1016/j.oret.2024.03.014
9. Cepeda, M. S., Katz, E. G., & **Blacketer, C**. (2017). Microbiome-Gut-Brain Axis: Probiotics and Their Association With Depression. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *29*(1), 39–44. https://doi.org/10.1176/appi.neuropsych.15120410

10. Cepeda, M. S., Kern, D. M., **Blacketer, C**., & Drevets, W. C. (2020). Low levels of cholesterol and the cholesterol type are not associated with depression: Results of a cross-sectional NHANES study. *Journal of Clinical Lipidology*, *14*(4), 515–521. https://doi.org/10.1016/j.jacl.2020.06.001
11. Cepeda, M. S., Reps, J., Fife, D., **Blacketer, C**., Stang, P., & Ryan, P. (2018). Finding treatment-resistant depression in real-world data: How a data- driven approach compares with expert-based heuristics. *Depression and Anxiety*, *35*(3), 220–228. https://doi.org/10.1002/da.22705
12. Cepeda, M. S., Stang, P., **Blacketer, C**., Kent, J. M., & Wittenberg, G. M. (2016). Clinical Relevance of Sleep Duration: Results from a Cross-Sectional Analysis Using NHANES. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *12*(6), 813–819. https://doi.org/10.5664/jcsm.5876
13. Cepeda, M. S., Weinstein, R., **Blacketer, C**., & Lynch, M. C. (2017). Association of flossing/inter-dental cleaning and periodontitis in adults. *Journal of Clinical Periodontology*, *44*(9), 866–871. https://doi.org/10.1111/jcpe.12765
14. Duarte-Salles, T., Vizcaya, D., Pistillo, A., Casajust, P., Sena, A. G., Lai, L. Y. H., Prats-Uribe, A., Ahmed, W.-U.-R., Alshammari, T. M., Alghoul, H., Alser, O., Burn, E., You, S. C., Areia, C., **Blacketer, C**., DuVall, S., Falconer, T., Fernandez-Bertolin, S., Fortin, S., … Prieto-Alhambra, D. (2020). Baseline characteristics, management, and outcomes of 55,270 children and adolescents diagnosed with COVID-19 and 1,952,693 with influenza in France, Germany, Spain, South Korea and the United States: An international network cohort study. *medRxiv : The Preprint Server for Health Sciences*, 2020.10.29.20222083. https://doi.org/10.1101/2020.10.29.20222083
15. Duarte-Salles, T., Vizcaya, D., Pistillo, A., Casajust, P., Sena, A. G., Lai, L. Y. H., Prats-Uribe, A., Ahmed, W.-U.-R., Alshammari, T. M., Alghoul, H., Alser, O., Burn, E., You, S. C., Areia, C., **Blacketer, C**., DuVall, S., Falconer, T., Fernandez-Bertolin, S., Fortin, S., … Prieto-Alhambra, D. (2021). Thirty-Day Outcomes of Children and Adolescents With COVID-19: An International Experience. *Pediatrics*, *148*(3), e2020042929. https://doi.org/10.1542/peds.2020-042929
16. Fife, D., **Blacketer, C**., Knight, K., & Weaver, J. (2021a). Stroke Risk Among Non-Elderly Users of Haloperidol or First-Generation Antipsychotics vs Second-Generation Antipsychotics: A Cohort Study from a US Health Insurance Claims Database. *Drugs - Real World Outcomes*, *8*(4), 481–496. https://doi.org/10.1007/s40801-021-00267-2
17. Fife, D., **Blacketer, C**., Knight, R. K., & Weaver, J. (2021b). Stroke Risk Among Elderly Users of Haloperidol and Typical Antipsychotics Versus Atypical Antipsychotics: A Real-World Study From a US Health Insurance Claims Database. *The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry*, *29*(5), 499–510. https://doi.org/10.1016/j.jagp.2020.09.017
18. Fife, D., **Blacketer, C**., Reps, J. M., & Ryan, P. (2018). Database Studies of Treatment-Resistant Depression Should Take Account of Adequate Dosing. *The Primary Care Companion for CNS Disorders*, *20*(4), 18m02274. https://doi.org/10.4088/PCC.18m02274

19. Fife, D., Reps, J., Cepeda, M. S., Stang, P., **Blacketer, C**., & Singh, J. (2018). Treatment resistant depression incidence estimates from studies of health insurance databases depend strongly on the details of the operating definition. *Heliyon*, *4*(7), e00707. https://doi.org/10.1016/j.heliyon.2018.e00707
20. Golozar, A., Lai, L. Y., Sena, A. G., Vizcaya, D., Schilling, L. M., Huser, V., Nyberg, F., Duvall, S. L., Morales, D. R., Alshammari, T. M., Abedtash, H., Ahmed, W.-U.-R., Alser, O., Alghoul, H., Zhang, Y., Gong, M., Guan, Y., Areia, C., Jonnagaddala, J., … **Blacketer, C.** … Prieto-Alhambra, D. (2020). Baseline phenotype and 30-day outcomes of people tested for COVID-19: An international network cohort including >3.32 million people tested with real-time PCR and >219,000 tested positive for SARS-CoV-2 in South Korea, Spain and the United States. *medRxiv : The Preprint Server for Health Sciences*, 2020.10.25.20218875. https://doi.org/10.1101/2020.10.25.20218875
21. Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., Payne, P. R. O., Pfaff, E. R., Robinson, P. N., Saltz, J. H., Spratt, H., Suver, C., Wilbanks, J., Wilcox, A. B., Williams, A. E., Wu, C., Blacketer, C., Bradford, R. L., Cimino, J. J., … **Blacketer, C**. … Gersing, K. R. (2021). The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association : JAMIA*, *28*(3), 427–443. https://doi.org/10.1093/jamia/ocaa196
22. Hufstedler, H., Roell, Y., Pena, A., Krishnan, A., Green, I., Barbosa-Silva, A., Kremer, A., **Blacketer, C**., Fortier, I., Howard, K., LeRoy, B., Hafeza, E., Baorto, D., Swertz, M., Maxwell, L., & Jaenisch, T. (2024). Navigating data standards in public health: A brief report from a data- standards meeting. *Journal of Global Health*, *14*, 03024. https://doi.org/10.7189/jogh.14.03024
23. Keloth, V. K., Banda, J. M., Gurley, M., Heider, P. M., Kennedy, G., Liu, H., Liu, F., Miller, T., Natarajan, K., V Patterson, O., Peng, Y., Raja, K., Reeves, R. M., Rouhizadeh, M., Shi, J., Wang, X., Wang, Y., Wei, W.-Q., Williams, A. E., … **Blacketer, C**. … Xu, H. (2023). Representing and utilizing clinical textual data for real world studies: An OHDSI approach. *Journal of Biomedical Informatics*, *142*, 104343. https://doi.org/10.1016/j.jbi.2023.104343
24. Khalid, S., Yang, C., **Blacketer, C**., Duarte-Salles, T., Fernandez-Bertolin, S., Kim, C., Park, R. W., Park, J., Schuemie, M. J., Sena, A. G., Suchard, M. A., You, S. C., Rijnbeek, P. R., & Reps, J. M. (2021). A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. *Computer Methods and Programs in Biomedicine*, *211*, 106394. https://doi.org/10.1016/j.cmpb.2021.106394
25. Khera, R., Aminorroaya, A., Dhingra, L. S., Thangaraj, P. M., Camargos, A. P., Bu, F., Ding, X., Nishimura, A., Anand, T. V., Arshad, F., **Blacketer, C**., Chai, Y., Chattopadhyay, S., Cook, M., Dorr, D. A., Duarte-Salles, T., DuVall, S. L., Falconer, T., French, T. E., … Suchard, M. A. (2024a). Comparative Effectiveness of Second-line Antihyperglycemic Agents for Cardiovascular Outcomes: A Large-scale, Multinational, Federated

Analysis of the LEGEND-T2DM Study. *medRxiv : The Preprint Server for Health Sciences*, 2024.02.05.24302354. https://doi.org/10.1101/2024.02.05.24302354

26. Khera, R., Aminorroaya, A., Dhingra, L. S., Thangaraj, P. M., Pedroso Camargos, A., Bu, F., Ding, X., Nishimura, A., Anand, T. V., Arshad, F., **Blacketer, C**., Chai, Y., Chattopadhyay, S., Cook, M., Dorr, D. A., Duarte-Salles, T., DuVall, S. L., Falconer, T., French, T. E., … Suchard, M. A. (2024b). Comparative Effectiveness of Second-Line Antihyperglycemic Agents for Cardiovascular Outcomes: A Multinational, Federated Analysis of LEGEND-T2DM. *Journal of the American College of Cardiology*, *84*(10), 904–917. https://doi.org/10.1016/j.jacc.2024.05.069
27. Khera, R., Dhingra, L. S., Aminorroaya, A., Li, K., Zhou, J. J., Arshad, F., Blacketer, C., Bowring, M. G., Bu, F., Cook, M., Dorr, D. A., Duarte-Salles, T., DuVall, S. L., Falconer, T., French, T. E., Hanchrow, E. E., Horban, S., Lau, W. C., Li, J., … **Blacketer, C**. … Suchard, M. A. (2023). Multinational patterns of second line antihyperglycaemic drug initiation across cardiovascular risk groups: Federated pharmacoepidemiological evaluation in LEGEND-T2DM. *BMJ Medicine*, *2*(1), e000651. https://doi.org/10.1136/bmjmed-2023-000651
28. Kim, C., Bu, F., **Blacketer, C**., Ostropolets, A., Duarte-Salles, T., Viernes, B., Falconer, T., Pistillo, A., Li, J., Yin, C., Van Zandt, M., Nagy, P., Nishimura, A., Minty, E., You, S. C., Sawano, M., Sawano, S., Jeon, J. Y., Aminorroaya, A., … Lu, Y. (2025). Comparative Safety of Second-Line Antihyperglycemic Agents in Older Adults with Type 2 Diabetes: A Multinational Real-World Evidence From LEGEND-T2DM. In *medRxiv: The preprint server for health sciences* (p. 2025.05.08.25327248). https://doi.org/10.1101/2025.05.08.25327248
29. Kostka, K., Duarte-Salles, T., Prats-Uribe, A., Sena, A. G., Pistillo, A., Khalid, S., Lai, L. Y. H., Golozar, A., Alshammari, T. M., Dawoud, D. M., Nyberg, F., Wilcox, A. B., Andryc, A., Williams, A., Ostropolets, A., Areia, C., Jung, C. Y., Harle, C. A., Reich, C. G., … **Blacketer, C**. … Prieto-Alhambra, D. (2022). Unraveling COVID-19: A Large-Scale Characterization of 4.5 Million COVID-19 Cases Using CHARYBDIS. *Clinical Epidemiology*, *14*, 369–384. https://doi.org/10.2147/CLEP.S323292
30. Krumme, A. A., Suruki, R. Y., **Blacketer, C**., Hardin, J., Swerdel, J. N., Tjoa, M. L., & Markham, K. B. (2025). Characterization of severity of hemolytic disease of the fetus and newborn due to Rhesus antigen alloimmunization. *AJOG Global Reports*, *5*(1), 100439. https://doi.org/10.1016/j.xagr.2024.100439
31. Morales, D. R., Bu, F., Viernes, B., DuVall, S. L., Matheny, M. E., Simon, K. R., Falconer, T., Richter, L. R., Ostropolets, A., Lau, W. C. Y., Man, K. K. C., Chattopadhyay, S., Mathioudakis, N., Minty, E., Nishimura, A., Sun, F., Yin, C., Seager, S. L., Chai, Y., … Suchard, M. A. (2025). Risk of Thyroid Tumors With GLP-1 Receptor Agonists: A Retrospective Cohort Study. *Diabetes Care*, *48*(8), 1386–1394. https://doi.org/10.2337/dc25-0154
32. Morales, D. R., Ostropolets, A., Lai, L., Sena, A., Duvall, S., Suchard, M., Verhamme, K., Rjinbeek, P., Posada, J., Ahmed, W., Alshammary, T., Alghoul, H., Alser, O., Areia, C., **Blacketer, C**., Burn, E., Casajust, P., You, S. C., Dawoud, D., … Kostka, K. (2023). Char-

acteristics and outcomes of COVID-19 patients with and without asthma from the United States, South Korea, and Europe. *The Journal of Asthma : Official Journal of the Association for the Care of Asthma*, *60*(1), 76–86. https://doi.org/10.1080/02770903.2021.2025392

33. Moreno-Martos, D., Verhamme, K., Ostropolets, A., Kostka, K., Duarte-Sales, T., Prieto-Alhambra, D., Alshammari, T. M., Alghoul, H., Ahmed, W.-U.-R., **Blacketer, C**., DuVall, S., Lai, L., Matheny, M., Nyberg, F., Posada, J., Rijnbeek, P., Spotnitz, M., Sena, A., Shah, N., … Morales, D. (2022). Characteristics and outcomes of COVID-19 patients with COPD from the United States, South Korea, and Europe. *Wellcome Open Research*, *7*, 22. https://doi.org/10.12688/wellcomeopenres.17403.3
34. Newbern, E. C., Shoaibi, A., Haynes, K., **Blacketer, C**., Willame, C., DeFalco, F., Rao, G. A., Davis, K., Velarde, L. A., Praet, N., Makadia, R., Xu, Y., Ryan, P., & Schuemie, M. (2025). A rapid cycle analytics framework for vaccine safety surveillance within a real-world data network: Experience with enhanced surveillance of the Janssen COVID-19 vaccine. *Vaccine*, *55*, 127044. https://doi.org/10.1016/j.vaccine.2025.127044
35. Nishimura, A., Xie, J., Kostka, K., Duarte-Salles, T., Bertolin, S. F., Aragon, M., **Blacketer, C**., Shoaibi, A., DuVall, S. L., Lynch, K., Matheny, M. E., Falconer, T., Morales, D. R., Conover, M. M., You, S. C., Pratt, N., Weaver, J., Sena, A. G., Schuemie, M. J., … Suchard, M. A. (2021). Alpha-1 blockers and susceptibility to COVID-19 in benign prostate hyperplasia patients: An international cohort study. *medRxiv : The Preprint Server for Health Sciences*, 2021.03.18.21253778. https://doi.org/10.1101/2021.03.18.21253778
36. Nishimura, A., Xie, J., Kostka, K., Duarte-Salles, T., Fernandez Bertolin, S., Aragon, M., **Blacketer, C**., Shoaibi, A., DuVall, S. L., Lynch, K., Matheny, M. E., Falconer, T., Morales, D. R., Conover, M. M., Chan You, S., Pratt, N., Weaver, J., Sena, A. G., Schuemie, M. J., … Suchard, M. A. (2022). International cohort study indicates no association between alpha-1 blockers and susceptibility to COVID-19 in benign prostatic hyperplasia patients. *Frontiers in Pharmacology*, *13*, 945592. https://doi.org/10.3389/fphar.2022.945592
37. Prieto-Alhambra, D., Kostka, K., Duarte-Salles, T., Prats-Uribe, A., Sena, A., Pistillo, A., Khalid, S., Lai, L., Golozar, A., Alshammari, T. M., Dawoud, D., Nyberg, F., Wilcox, A., Andryc, A., Williams, A., Ostropolets, A., Areia, C., Jung, C. Y., Harle, C., …**Blacketer, C**. … Suchard, M. (2021). Unraveling COVID-19: A large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. *Research Square*, rs.3.rs-279400. https://doi.org/10.21203/rs.3.rs-279400/v1
38. Rao, G. A., Shoaibi, A., Makadia, R., Hardin, J., Swerdel, J., Weaver, J., Voss, E. A., Conover, M. M., Fortin, S., Sena, A. G., Knoll, C., Hughes, N., Gilbert, J. P., **Blacketer, C**., Andryc, A., DeFalco, F., Molinaro, A., Reps, J., Schuemie, M. J., & Ryan, P. B. (2025). CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization. *PloS One*, *20*(1), e0310634. https://doi.org/10.1371/journal.pone.0310634
39. Reyes, C., Pistillo, A., Fernandez-Bertolin, S., Recalde, M., Roel, E., Puente, D., Sena, A. G., **Blacketer, C**., Lai, L., Alshammari, T. M., Ahmed, W.-U.-R., Alser, O., Alghoul, H., Areia, C.,

Dawoud, D., Prats-Uribe, A., Valveny, N., de Maeztu, G., Sorli Redo, L., … Duarte-Salles, T. (2021). Characteristics and outcomes of patients with COVID-19 with and without prevalent hypertension: A multinational cohort study. *BMJ Open*, *11*(12), e057632. https://doi.org/10.1136/bmjopen-2021-057632

40. Roel, E., Pistillo, A., Recalde, M., Sena, A. G., Fernandez-Bertolin, S., Aragon, M., Puente, D., Ahmed, W.-U.-R., Alghoul, H., Alser, O., Alshammari, T. M., Areia, C., **Blacketer,** C., Carter, W., Casajust, P., Culhane, A. C., Dawoud, D., DeFalco, F., DuVall, S. L., … Duarte-Salles, T. (2021). Characteristics and Outcomes of Over 300,000 Patients with COVID-19 and History of Cancer in the United States and Spain. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *30*(10), 1884–1894. https://doi.org/10.1158/1055-9965.EPI-21-0266
41. Shin, S. J., You, S. C., Park, Y. R., Roh, J., Kim, J.-H., Haam, S., Reich, C. G., **Blacketer, C**., Son, D.-S., Oh, S., & Park, R. W. (2019). Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *Journal of Medical Internet Research*, *21*(3), e13249. https://doi.org/10.2196/13249
42. Shoaibi, A., Rao, G. A., Voss, E. A., Ostropolets, A., Mayer, M. A., Ramirez-Anguita, J. M., Maljkovic, F., Carevic, B., Horban, S., Morales, D. R., Duarte-Salles, T., Fraboulet, C., Le Carrour, T., Denaxas, S., Papez, V., John, L. H., Rijneek, P. R., Minty, E., Alshammari, T. M., …**Blacketer, C**.… Ryan, P. B. (2022). Phenotype Algorithms for the Identification and Characterization of Vaccine-Induced Thrombotic Thrombocytopenia in Real World Data: A Multinational Network Cohort Study. *Drug Safety*, *45*(6), 685–698. https://doi.org/10.1007/s40264-022-01187-y
43. Tan, E. H., Sena, A. G., Prats-Uribe, A., You, S. C., Ahmed, W.-U.-R., Kostka, K., Reich, C., Duvall, S. L., Lynch, K. E., Matheny, M. E., Duarte-Salles, T., Bertolin, S. F., Hripcsak, G., Natarajan, K., Falconer, T., Spotnitz, M., Ostropolets, A., **Blacketer, C**., Alshammari, T. M., … Prieto-Alhambra, D. (2020). Characteristics, outcomes, and mortality amongst 133,589 patients with prevalent autoimmune diseases diagnosed with, and 48,418 hospitalised for COVID-19: A multinational distributed network cohort analysis. *medRxiv : The Preprint Server for Health Sciences*, 2020.11.24.20236802. https://doi.org/10.1101/2020.11.24.20236802
44. Tan, E. H., Sena, A. G., Prats-Uribe, A., You, S. C., Ahmed, W.-U.-R., Kostka, K., Reich, C., Duvall, S. L., Lynch, K. E., Matheny, M. E., Duarte-Salles, T., Bertolin, S. F., Hripcsak, G., Natarajan, K., Falconer, T., Spotnitz, M., Ostropolets, A., **Blacketer, C**., Alshammari, T. M., … Prieto-Alhambra, D. (2021). COVID-19 in patients with autoimmune diseases: Characteristics and outcomes in a multinational network of cohorts across three countries. *Rheumatology (Oxford, England)*, *60*(SI), SI37–SI50. https://doi.org/10.1093/rheumatology/keab250
45. Voss, E. A., **Blacketer, C**., van Sandijk, S., Moinat, M., Kallfelz, M., van Speybroeck, M., Prieto-Alhambra, D., Schuemie, M. J., & Rijnbeek, P. R. (2023). European Health Data & Evidence Network-learnings from building out a standardized international health

data network. *Journal of the American Medical Informatics Association : JAMIA*, *31*(1), 209–219. https://doi.org/10.1093/jamia/ocad214

46. Voss, E. A., Shoaibi, A., Yin Hui Lai, L., **Blacketer, C**., Alshammari, T., Makadia, R., Haynes, K., Sena, A. G., Rao, G., van Sandijk, S., Fraboulet, C., Boyer, L., Le Carrour, T., Horban, S., Morales, D. R., Martinez Roldan, J., Ramirez-Anguita, J. M., Mayer, M. A., de Wilde, M., … Ryan, P. B. (2023). Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 databases: A multinational retrospective cohort study. *EClinicalMedicine*, *58*, 101932. https://doi.org/10.1016/j.eclinm.2023.101932
47. Wang, Y., Hester, L. L., Lofland, J., Rose, S., Karyekar, C. S., Kern, D. M., **Blacketer, C**., Davis, K., & Shields-Tuttle, K. (2022). Update on prevalence of diagnosed systemic lupus erythematosus (SLE) by major health insurance types in the US in 2016. *BMC Research Notes*, *15*(1), 5. https://doi.org/10.1186/s13104-021-05877-1

ABOUT THE AUTHOR



Margaret Sinclair Blacketer (Clair) was born in 1988 in Chesapeake, Virginia. She attended Princess Anne High School where she earned an International Baccalaureate Diploma. She went on to attend James Madison University to study biology, planning to go into medicine. Following a realization that medical school was not in the cards, Clair quickly pivoted to epidemiology. With that in mind, she pursued a master's degree in public health from Eastern Virginia Medical School where she fell in love with data science.

While in grad school, Clair found a position as a clinical research assistant at a local diabetes center where she worked on clinical trials. After graduation, she went on to a Virginia hospital system performing investigator-initiated studies using data from 12 hospitals. Clair then went on to a large health insurance company before joining the Global Epidemiology Organization at Johnson & Johnson in 2015.

Clair currently lives in Virginia Beach with her husband, Joey, son James, two dogs Mayonnaise and Sandwich, and Poppy the crazy cat.