Generating Reliable and Responsive Observational Evidence:

Reducing Pre-analysis Bias

Anna Ostropolets

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Generating Reliable and Responsive Observational Evidence:

Reducing Pre-analysis Bias

Anna Ostropolets

A growing body of evidence generated from observational data has demonstrated the potential to influence decision-making and improve patient outcomes. For observational evidence to be actionable, however, it must be generated reliably and in a timely manner. Large distributed observational data networks enable research on diverse patient populations at scale and develop new sound methods to improve reproducibility and robustness of real-world evidence. Nevertheless, the problems of generalizability, portability and scalability persist and compound. As analytical methods only partially address bias, reliable observational research (especially in networks) must address the bias at the design stage (i.e., pre-analysis bias) including the strategies for identifying patients of interest and defining comparators.

This thesis synthesizes and enumerates a set of challenges to addressing pre-analysis bias in observational studies and presents mixed-methods approaches and informatics solutions for overcoming a number of those obstacles. We develop frameworks, methods and tools for scalable and reliable phenotyping including data source granularity estimation, comprehensive concept set selection, index date specification, and structured data-based patient review for phenotype evaluation. We cover the research on potential bias in the unexposed comparator definition including systematic background rates estimation and interpretation, and definition and evaluation of the unexposed comparator.

We propose that the use of standardized approaches and methods as described in this thesis not only improves reliability but also increases responsiveness of observational evidence. To test this hypothesis, we designed and piloted a Data Consult Service - a service that generates new on-demand evidence at the bedside. We demonstrate that it is feasible to generate reliable evidence to address clinicians' information needs in a robust and timely fashion and provide our analysis of the current limitations and future steps needed to scale such a service.

Table of Contents

List of Abbreviations

AU-ePBRN: Australian Electronic Practice-Based Research Network

CCAE: Commercial Claims and Encounters Database

CDISC SDTM: Clinical Data Interchange Standards Consortium Study Data Tabulation Model

CDM: Common Data Model

CKD: chronic kidney disease

COPD: chronic obstructive lung disorder

CPRD: Clinical Practice Research Datalink

CPT4: Current Procedural Terminology, version 4

CUIMC: Columbia University Irving Medical Center

DA: Disease Analyzer

eGFR: estimated glomerular filtration rate

EHR: Electronic Health Record

eMERGE: Electronic Medical Records and Genomics

ESRD: end-stage renal disease

FDA: Food and Drug Administration

GLP1-RA: glucagon-like peptide 1 receptor agonists

HCPCS: Healthcare Common Procedure Coding System

HCUP: Healthcare Cost and Utilization Nationwide database

i2b2: Integrating Biology & the Bedside

ICD-10(CM): International Classification of Diseases, Tenth Revision, Clinical Modification

ICD-9(CM): International Classification of Diseases, Ninth Revision, Clinical Modification

IPCI: Integrated Primary Care Information database

IQR: Interquartile range

IR: Incidence rate

IRR: Incidence rate ratio

JMDC: Japan Medical Data Center

LOINC: Logical Observation Identifiers Names and Codes

LPD: Longitudinal Patient Data

MDCD: MarketScan Multi-State Medicaid Database

MDCR: MarketScan Medicare Supplemental Database

MIMIC-III: Medical Information Mart for Intensive Care, version III

NDC: National Drug Codes

NHIS: National Health Insurance Service

NSC: National Sample Cohort

OHDSI: Observational Health Data Sciences and Informatics

OMOP: Observational Medical Outcomes Partnership

PHD: Premier Healthcare Database

PPV: positive predictive value

SD: Standard deviation

SDM: Standardized difference of means

SNOMED-CT: Systematized Nomenclature of Medicine -- Clinical Terms

STaRR: Stanford Medicine Research Data Repository

T1DM: type II diabetes mellitus

T2DM: type I diabetes mellitus

List of Figures

List of Tables

List of tables

Glossary

Cohort is the set of persons satisfying one or more inclusion and exclusion criteria for a duration of time.

Cohort definition is operationalized executable algorithm that translates a set of inclusion and exclusion criteria, which, when executed against a data source, produces a cohort of patients.
In this thesis we use the terms cohort definition and phenotype definition interchangeably.

Cohort start and end date are the points in time the subjects enter and exit the cohort respectively.

Concept (code, ontology term) - a unit of symbolic processing in controlled ontologies, terminologies, and vocabularies, which embodies a particular meaning [1].

Concept set (code set, value set) is the set of medical terms used to represent the data elements in phenotype definition [2].

Interoperability is the ability of different information systems, devices, and applications (systems) to access, exchange, integrate, and cooperatively use data in a coordinated manner, within and across organizational, regional, and national boundaries, to provide timely and seamless portability of information and optimize the health of individuals and populations [3].

Portability is an ability of phenotype to be implemented faithfully and easily at a different site while maintaining a similar performance or, in other terms, performance over time and effort [4,5].

Phenotype is a specification of an observable, potentially changing state of an organism, as distinguished from the genotype, which is derived from an organism's genetic makeup [4].

Representation is the organization, structuring, or labeling of data, information, or knowledge in a formalized manner intended to support subsequent processing (comprehension, decisions, and actions) and optimize communication [6]

Publications

The work conducted in this thesis led to the following first or co-first (*) author publications:

1. ***Ostropolets A***, Reich C, Ryan P, Weng C, Molinaro A, DeFalco F, *et al*. Characterizing database granularity using SNOMED-CT hierarchy. *AMIA Annual Symposium proceedings* (2020) [Second Place, AMIA Student Paper Competition 2020] [7]

2. ***Ostropolets A***, Ryan P, Hripcsak G. Phenotyping in distributed data networks: selecting the right codes for the right patients. *AMIA Annual Symposium proceedings* (2022) [8]

3. ***Ostropolets A***, Reich C, Ryan P, Shang N, Hripcsak G, Weng C. Adapting electronic health records-derived phenotypes to claims data: Lessons learned in using limited clinical data for phenotyping. *Journal of Biomedical Informatics* (2020) [9]

4. ***Ostropolets A***, Li X, Makadia R, Rao G, Rijnbeek PR, Duarte-Salles T, et al. Factors Influencing Background Incidence Rate Calculation: Systematic Empirical Evaluation Across an International Network of Observational Databases. *Frontiers in Pharmacology* (2022) [10]

5. ***Ostropolets A***, Ryan PB, Schuemie MJ, Hripcsak G. Characterizing Anchoring Bias in Vaccine Comparator Selection Due to Health Care Utilization With COVID-19 and Influenza: Observational Cohort Study. *JMIR Public Health Surveillance* (2022) [Best Community Contribution, OHDSI 2021] [11]

6. ***Ostropolets A***, Hripcsak G. COVID-19 vaccination effectiveness rates by week and sources of bias: a retrospective cohort study. *British Medical Journal Open* (2022) [12]

7. ***Ostropolets A***, Chen R, Zhang L, Hripcsak G. Characterizing physicians' information needs related to a gap in knowledge unmet by current evidence. *Journal of the American Medical Informatics Association Open* (2020) [13]

8. ***Ostropolets A***, Zhang L, Hripcsak G. A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. *Journal of the American Medical Informatics Association* (2020) [14]

9. ***Ostropolets A***, Zachariah P, Ryan P, Chen R, Hripcsak G. Data Consult Service: Can we use observational data to address immediate clinical needs? *Journal of the American Medical Informatics Association* (2021) [15]

10. Li X, ***Ostropolets A*,*** Makadia R, Shoaibi A, Rao G, Sena AG, et al. Characterising the background incidence rates of adverse events of special interest for COVID-19 vaccines in eight countries: multinational network cohort study. *British Medical Journal* (2021) [16]

11. ***Ostropolets A***, Elias PA, Reyes MV, Wan EY, Pajvani UB, Hripcsak G, et al. Metformin Is Associated with a Lower Risk of Atrial Fibrillation and Ventricular Arrhythmias Compared to Sulfonylureas: An Observational Study. *Circulation: Arrhythmia and Electrophysiology* (2021) [17]

Other co-author publications included in the thesis:

12. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MFT, … ***Ostropolets A***, … *et al*. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nature communications* (2020) [18]

13. Castano VG, Spotnitz M, Waldman GJ, Joiner EF, Choi H, ***Ostropolets A***, *et al*. Identification of patients with drug resistant epilepsy in electronic medical record data using the Observational Medical Outcomes Partnership Common Data Model. *Epilepsia* (2022) [19]

14. Chen R, Schuemie MJ, Suchard MA, ***Ostropolets A***, Zhang L, Ryan PB *et al*. Evaluation of Large-scale Propensity Score Modeling and Covariate Balance on Potential Unmeasured Confounding in Observational Research. *AMIA Annual Symposium proceedings* (2020) [20]

15. Khera R, Schuemie MJ, Lu Y, ***Ostropolets A***, Chen RJ, G Hripcsak G, *et al*. Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies. *BMJ Open* (2022) [21]

16. Kostka K, Duarte-Salles T, Prats-Uribe A, Sena AG, Pistillo A, Khalid S,… ***Ostropolets A***, … *et al*. Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. *Clinical Epidemiology* (2022) [22]

17. Lane JC, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, … ***Ostropolets A***, … *et al*. Risk of depression, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multinational network cohort study. *Rheumatology* (2021) [23]

18. Lane JC, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, … ***Ostropolets A***,… *et al*. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. *The Lancet Rheumatology* (2020) [24]

19. Li X, Lai LY, ***Ostropolets A***, Arshad F, Tan EH, Casajust P, *et al*. Bias, precision and timeliness of historical (background) rate comparison methods for vaccine safety monitoring: an empirical multi-database analysis. *Frontiers in Pharmacology* (2021) [25]

20. Morales DR, ***Ostropolets A***, Lai L, Sena A, Duvall S, Suchard M, *et al*. Characteristics and outcomes of COVID-19 patients with and without asthma from the United States, South Korea, and Europe. *Journal of Asthma* (2022) [26]

21. Moreno-Martos D, Verhamme K, ***Ostropolets A***, Kostka K, Duarte-Sales T, *et al*. Characteristics and outcomes of COVID-19 patients with COPD from the United States, South Korea, and Europe. *Wellcome Open Research* (2022) [27]

22. Reps JM, Kim C, Williams RD, Markus AF, Yang C, … ***Ostropolets*** A,… *et al*. Implementation of the COVID-19 vulnerability index across an international network of health care data sets: collaborative external validation study. *JMIR medical informatics* (2021) [28]

23. Reyes C, Pistillo A, Fernández-Bertolín S, Recalde M, Roel E, Puente D, ... ***Ostropolets A***, …*et al*. Characteristics and outcomes of patients with COVID-19 with and without prevalent hypertension: a multinational cohort study. *BMJ Open* (2021) [29]

24. Rodriguez VA, Tony S, Thangaraj P, Pang C, Kalluri KS, Jiang X, ***Ostropolets A***, *et al*. Phenotype Concept Set Construction from Concept Pair Likelihoods. *AMIA Annual Symposium Proceedings* (2020) [30]

25. Roel E, Pistillo A, Recalde M, Sena AG, Fernández-Bertolín S, Aragón M, ... ***Ostropolets A***, … *et al*. 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* (2021) [31]

26. Seong Y, You SC, ***Ostropolets A***, *et al*. Incorporation of Korean Electronic data interchange vocabulary into observational medical outcomes partnership vocabulary. *Healthcare Informatics Research* (2021) [32]

27. Shoaibi A, Rao GA, Voss EA, ***Ostropolets A***, Mayer MA, *et al*. Phenotype Algorithms for the Identification and Characterization of Vaccine-Induced Thrombotic Thrombocytopenia in Real World Data: A Multinational Network Cohort Study. *Drug Safety* (2022) [33]

28. Spotnitz M, ***Ostropolets A***, Castano VG, Natarajan K, Waldman GJ, *et al*. Patient characteristics and antiseizure medication pathways in newly diagnosed epilepsy: Feasibility and pilot results using the common data model in a single-center electronic medical record database. *Epilepsy & Behavior* (2022) [34]

29. Tan EH, Sena AG, Prats-Uribe A, You SC, Ahmed WUR, Kostka K, … ***Ostropolets A***,… *et al*. COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries. *Rheumatology* (2021) [35]

Acknowledgments

This dissertation would not have been possible without the contributions of many people. First and foremost, I would like to express my gratitude to my advisor, George Hripcsak, for his continued guidance and support over the last four years. His scientific rigor, way of thinking about research problems and work ethics shaped me as a researcher. His and Patrick Ryan's mentorship made me a better scientist, a better colleague, and a better person.

I sincerely appreciate Chunhua Weng, Hongfang Liu and Harlan Krumholz for serving on my dissertation committee and devoting their time, and commitment to improving my dissertation and research.

This work could not exist without the fruitful and welcoming OHDSI consortium, which has grown to an incredible community that shapes the way observational research is conducted and viewed. And we continue to grow as the readers go through this thesis.

I would like to thank my friends at DBMI and outside of it for your support. And special thanks to Harry Reyes, who has been a continuous source of ideas, methods advice, and encouragement.

And, of course, my family. I hope that I made these 4 years away from you count. Your impact on me and my research career cannot be overestimated.

Chapter 1. Introduction

1.1. Problem statement: need for reliable and responsive evidence

Accumulated scientific evidence informs clinical decisions to promote better care, improve patient outcomes, and reduce costs [36–41]. Randomized clinical trials (RCTs) are the backbone of medical evidence and the gold standard for inferring intervention benefits and risks. Nevertheless, they are often not generalizable to real-world patients [42–46], expensive, slow and require post-marketing studies for new drugs [47,48], which, in turn, only delays evidence delivery.

As a result, many clinical questions remain unanswered [13,43,49–52], which produces variability in clinical practice [53], especially evident as new diseases emerge. For example, the COVID-19 pandemic illustrated that a lack of evidence leads to the use of treatment regimens of unknown effectiveness, off-label drug use, and suboptimal patient care [54,55].

A growing body of observational data, such as electronic health records (EHRs) and administrative claims, could be a solution to the lack of responsive evidence. These data sources have been continuously used in observational drug effectiveness and surveillance studies, precision medicine initiatives, prediction tools, and other clinical decision support systems (CDSS) and have shown the potential to influence decision-making [56,57]. When individual data sources do not meet sample-size requirements to support valid inference, large distributed observational data networks enable clinical research on diverse patient populations at scale, bringing us closer to a dream of providing accurate and reliable treatment recommendations to the patients the moment they seek care.

Reliability of observational evidence is commonly criticized, with analytical approaches and methods only partially addressing potential bias. Current research addresses the questions of reproducibility, replicability, transparency, and robustness of real-world evidence, with problems of generalizability and scalability remaining largely unsolved [57–62]. Bias can stem from inaccurate measurement of exposure, outcome or covariates or inaccurate comparator definition, leading to faulty evidence [63]. While there is a solid body of research on different sources of bias and its implication [4,64–71], there is a lack of readily available informatics solutions to systematically and reliably assess bias at the pre-analysis (design) stage, especially when operating heterogeneous data from multiple institutions.

The lack of standardized and scalable informatics solutions compromises timeliness. Even after a decade of observational research in networks, months and years are spent on designing observational studies, ensuring data plausibility and quality, and developing and validating phenotyping algorithms [72,73]. Despite the efforts to standardize these processes, they remain variable, labor-intense, and time-consuming, jeopardizing both reliability and timeliness [74–77]

1.2. Thesis approach and contribution

In this thesis, we hypothesize that actionable observational evidence can be generated in near-real time to reliably address clinicians' immediate information needs not covered by the existing evidence.

To achieve this, we develop mixed-method multidisciplinary approaches and informatics tools to systematically assess and mitigate bias at the pre-analytical stage of effectiveness and safety studies. We propose and subsequently test the hypothesis that building scalable and robust pipelines to address bias in networks enables both more robust and faster evidence generation.

In this thesis, we systematically develop methods to investigate and mitigate bias at each step of a typical observational study from (a) creating and evaluating phenotyping algorithms for identifying patients of interest through (b) assessing and mitigating pre-analysis bias in comparator definition (temporal, selection, and other) to (c) generating and delivering evidence to clinicians at Columbia University Irving Medical Center (CUIMC) and NewYork-Presbyterian (NYP) Hospital.

1.2.1 Overall contribution

This thesis provides conceptual, methodological, and empirical contributions to biomedical informatics (Figure 1, further discussed in Section 1.2.2).

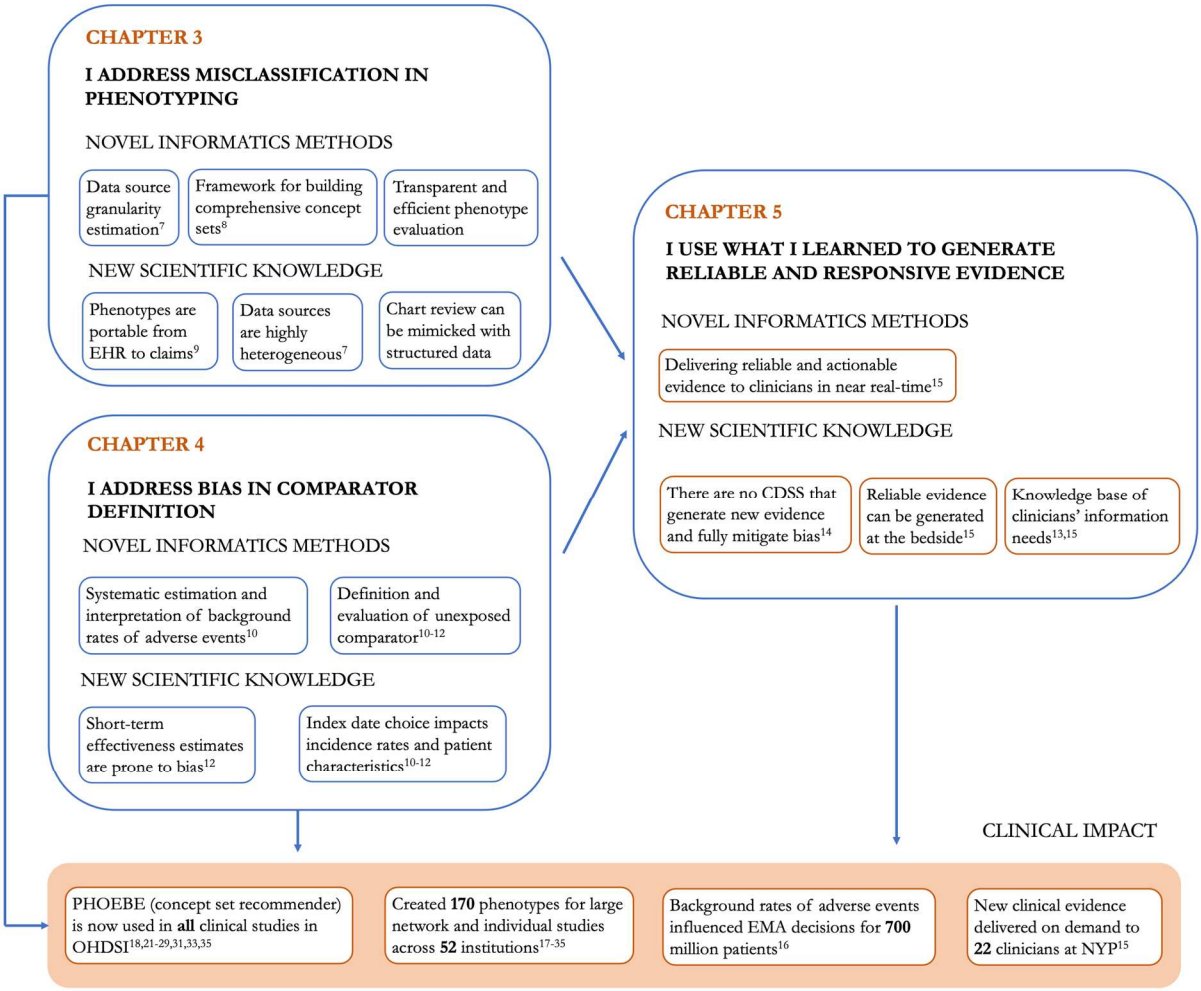


Figure 1. Summary of the contributions. NYP – NewYork-Presbyterian Hospital, EMA – European Medicines Agency, CDSS – clinical decision support system

Conceptual contribution

We developed novel frameworks for phenotype development and evaluation in distributed data networks, including frameworks for (a) building comprehensive concept sets and (b) reviewing structured patient data for case ascertainment.

These frameworks enable scalable and generalizable phenotyping both in networks and on individual data sources. The framework for building comprehensive concept sets informed our recommender system (PHOEBE) that helps clinicians create portable and generalizable concept sets without access to data. The framework for reviewing structured patient data informed a tool (KEEPER) that facilitates fast, transparent and reliable phenotype evaluation on EHR and claims data sources.

Methodological contribution

We developed original methods to identify and mitigate potential bias and measurement error in phenotyping and unexposed comparator definition, including methods for (a) data source granularity estimation, (b) systematic background incidence rates estimation and interpretation, and (c) definition and evaluation of an unexposed comparator.

Methods for granularity estimation enabled first large-scale characterization of 22 US and international data sources. Methods for background rate estimation and unexposed comparator definition provide informatics solutions for timely and reliable safety surveillance.

Empirical contribution

We identified major challenges related to scalable, reliable, and generalizable phenotyping in networks and characterized them at scale, which contributed to scientific knowledge.

This thesis had direct clinical impact through evidence delivery to clinicians, researchers and regulatory bodies. We (a) developed and disseminated 170 phenotype algorithms for a broad spectrum of clinical questions, (b) designed and deployed a service that generated new on-demand evidence to address 24 clinical questions for 22 clinicians at NewYork-Presbyterian Hospital and (c) generated background rates of adverse events of special interest, which impacted European Medicines Agency decisions regarding COVID-19 vaccine safety for more than 700 million patients.

1.2.2 Outline and detailed contribution

The background section (**Chapter 2**) presents an overview of the challenges in using observational data to generate reliable real-world evidence. It discusses bias in observational studies followed by an overview of challenges and advances of evidence generation in networks, accompanied by a more detailed analysis of gaps and related work in corresponding sections.

Algorithms for identifying patients of interest in observational data (phenotyping algorithms) are the key component of any observational study. Accurate phenotyping is critical for study validity, yet the challenge of creating scalable and efficient phenotypes remains largely unsolved. **Chapter 3** addresses phenotyping algorithm development and evaluation as a source of measurement error in distributed clinical networks. As phenotyping in networks is complicated by the fact that the data are collected across institutions from different countries capturing different aspects of care, we addressed the questions of phenotype portability and generalizability across disparate healthcare systems with the researchers not having the access to all data in the network.

In this chapter we systematically addressed challenges of rule-based and mixed-method phenotype development and evaluation starting with (a) selecting concepts that represent a clinical idea to (b) applying Boolean or temporal logic to create an operational definition to (c) assessing phenotype portability and performance.

First, in ***Section 3.1*** we investigate real-world concept (e.g., ICD-10 codes) utilization in a large international distributed data network and its impact on portability of phenotyping algorithms. We collect 272 billion records summarized from 22 international and US data sources across the Observational Health Data Sciences and Informatics (OHDSI) network and use them to study data source heterogeneity and granularity.

Using this data, we uncovered high data source heterogeneity and discrepancies in coding practices, which plays a crucial role in concept selection and, in turn, patient selection and composition. To identify the common patterns of code use we create a novel method for computing overall data source granularity. We then show three SNOMED-based approaches consistently classifying data source granularity, demonstrate common patterns in granularity based on the provenance of data and country of origin and analyze factors that influence granularity. This study is the first study to examine data source heterogeneity and characterize code utilization patterns at scale.

Then, in ***Section 3.2***, we use the same dataset to develop and evaluate a mixed-method recommender system for creating comprehensive concept sets to improve phenotype portability, scalability and efficiency. We evaluate its performance for four conditions (diabetes mellitus type I and II, attention deficit hyperactivity disorder and heart failure) on three EHR and claims data sources and show that it helps to identify more patients of interest, capture them early in the course of the disease and create concept sets that are portable across the network. As our recommender system (PHOEBE) enables large-scale phenotyping, it became a part of the standardized pipeline for phenotype development, evaluation and storage in the OHDSI network and has been used in individual studies as well as 11 major network studies.

Section 3.3 focuses on quantifying the variability in the operationalization of a conceptual definition into an executable algorithm. We conduct an experiment with 45 OHDSI collaborators set up as a standardized implementation of a clinical description from a recent high-impact observational study and analyze variation in implementations and its impact on patient population size and composition. We demonstrate that the small nuances of how the inclusion and exclusion criteria are implemented significantly influence patient composition even when isolated from other factors.

Subsequently, ***Section 3.4*** dives deeper into the complex problem of phenotype portability across different institutions. We inspect the factors that impact portability of EHR-derived phenotypes to claims data sources using chronic kidney disease as an example. We apply the insights learned in the previous sections and demonstrate the importance of data harmonization and concept set standardization in phenotype portability, including a more focused study of the influence of incorporating procedure codes into diagnosis-based phenotypes.

Finally, ***Section 3.5*** focuses on phenotype evaluation – one of the main bottlenecks of scalable phenotyping. The current gold standard – manual chart review – is interpretable and trustworthy but time-consuming, variable, and non-scalable. We propose Knowledge-Enhanced Electronic Profile Review system (KEEPER), a generalizable and scalable "chart review alternative" that represents patient structured data in a standardized way guided by the principles of clinical reasoning to ascertain individual patient status. We evaluate its utility and performance on four conditions (diabetes mellitus type 1, acute appendicitis, end stage renal disorder and chronic obstructive pulmonary disease) and demonstrate that, compared to manual chart review, its use achieves better consistency among reviewers and non-inferior accuracy in classifying patients at a fraction of time. KEEPER can enable reliable and scalable phenotyping on EHR and claims data sources.

Chapter 4 describes informatics solutions to temporal, selection, anchoring, and healthcare seeking-related biases in comparator definition and centers around safety (***Section 4.1***) and effectiveness (***Section 4.2***) studies. Selecting an appropriate comparator is challenging as the comparator should serve as a proxy for a counterfactual of the exposed population — what would have happened to those same individuals had they not been exposed — and any deviation between the comparator and that counterfactual represents a potential bias. Having an unexposed group as the comparator is especially

hard as the unexposed group generally represents a more heterogeneous population and does not have a clear disease onset date or exposure start date, deviating more from the exposed population. In this chapter, we investigate unexposed comparator selection strategies and their influence on baseline patient characteristics, background incidence rates (typically computed for observed-to-expected analysis common in drug surveillance) and vaccine effectiveness estimates. We use COVID-19 vaccination as an example of the target exposure given its public health importance.

In ***Section 4.1***, we conduct a systematic experiment on 12 US and international data sources, examining the influence of population characteristics, temporal trends, time-at-risk and index date choices on background rates of 15 adverse events of special interest. We observe that the background rates vary up to a factor of 1,000 across age groups, even after adjusting for age and sex and are highly sensitive to the choice of the index date (a visit, an arbitrary date or vaccination), data source, clean window choice and time-at-risk duration. Given these findings, we provide recommendations for (a) interpreting background rates within the context of study parameter choices and (b) improving robustness of observed-to-expected analysis. Along with creating novel methods for contextualizing background rates, we calculate the background rates themselves, which are later used by the European Medicines Agency to assess COVID-19 vaccine safety for more than 700 million people.

Given that during the COVID-19 pandemic the researchers have relied on methods developed for influenza vaccines, we additionally evaluate the impact of index date choice on patient characteristics for influenza and COVID-19 vaccines. We study two alternative selection procedures (anchoring) for the index date in the unexposed group based on how vaccines are administered - coupled or decoupled to another healthcare encounter. Using the data from two EHR data sources we show that anchoring

influences patient baseline characteristics and provide recommendations for empirical selection of anchors.

In ***Section 4.2***, we dive deeper into the bias in vaccine effectiveness studies. Given mixed reports regarding the effectiveness of COVID-19 vaccines during the first two weeks after the first dose, we investigate both short-term and long-term COVID-19 vaccine effectiveness, we conduct an analysis of short-term and long-term COVID-19 vaccine effectiveness accompanied by several secondary analyses and chart review to discover and mitigate selection and health-seeking behavior biases as well as confounding by severity and indication. We show that in a short time-at-risk interval, robust methods like large-scale propensity score matching and negative controls may fail to adjust for biases stemming from different health-seeking behavior in vaccinated and unvaccinated groups. The analysis is accompanied by the recommendations for vaccine effectiveness result reporting.

Chapter 5 showcases the use of methods and practices to mitigate pre-analysis bias when generating and delivering new evidence at the bedside and is centered around Data Consult Service – a system that uses observational data to generate new evidence at NewYork-Presbyterian Hospital.

In ***Section 5.1*** we conduct interviews with 31 clinician at Columbia University Irving Medical Center and use thematic analysis to explore their unmet evidence-related needs, which results in a modern taxonomy of needs not covered by the current guidelines and further informs the target groups and use cases for our service. We demonstrate that despite the abundance of knowledge, clinicians have multiple questions that are not covered by the current evidence and oftentimes have shared areas of unmet needs such as optimal treatment for patients with multiple comorbidities, patients with rare disorders, children, or those taking new drugs. Along with gathering the potential topics for our service, we identify potential target groups (experienced and inpatient physicians).

We then conduct a scoping review (***Section 5.2***) to investigate existing clinical decision support tools that aim at covering this gap to learn from their bias-mitigating strategies and features. We observe that the tools oftentimes lack demonstration of their utility and their impact on healthcare processes and patient outcomes remains unclear. Moreover, only one system attempts to properly address bias, highlighting a need for robust and reliable evidence generation pipelines.

Section 5.3 describes the process of designing and piloting a Data Consult Service at NYP. We implement a pipeline (question gathering, data exploration, iterative patient phenotyping, study execution, and evidence validity assessment) for generating new evidence in near-real time, which results in 24 answered questions collected from 22 clinicians. The answers satisfy clinicians information needs and they are willing to share the new evidence with their colleagues. We identify the key components required for successful early-stage implementation such as proactive involvement of the study team and participation in clinical rounds and shadowing. We classify and describe in-depth the main challenges in timely evidence delivery such as missing and incomplete data, underreported conditions, nonspecific coding and accurate identification of drug regimens.

Finally, **Chapter 6** summarizes the content of this thesis and outlines directions for future work.

Chapter 2. Background and related work

In this section, we will discuss how observational data can be used in real-world evidence (RWE) generation to complement evidence from randomized experiments. We will first talk about the common threats to validity that undermine the credibility of RWE. We will then articulate the advances in observational networks aimed at improving reliability and scalability of RWE followed by a brief introduction to one of the large observational data networks – OHDSI.

2.1 Observational data and threats to validity

Since the 1990s, when the concept of evidence-based medicine was first introduced, it has become the leading paradigm in clinical practice, shaping the way we view medicine today. Accumulated scientific evidence informs clinical decisions and policies to promote better care, improve patient outcomes, and reduce costs [36–41].

Nevertheless, medical evidence is neither comprehensive nor precise. Randomized clinical trials (RCTs), which are considered to be the gold standard for causal inference, have pitfalls [78]. RCTs are oftentimes not generalizable to real-world patients [42–44] and consider only a subset of the population, excluding patients with advanced cancer, chronic kidney disorder or systemic disorders, the elderly, pregnant women, and other vulnerable populations [79,80]. Clinical trials tend to focus narrowly on one condition at a time, which rarely provides clear guidance for patients with multiple conditions [45,80] or those undergoing complex interventions [46,81,82]. Trials usually include only a small number of patients, and are expensive and slow. After they are completed, post-marketing safety studies are required for new drugs [47,48], which, in turn, only delays evidence delivery.

Along with the continuing accumulation of practice-based evidence, observational data has made new approaches to evidence generation available [38,56,83]. The large body of data collected for clinical care and billing purposes can be repurposed to generate new evidence and studies on such data sources (EHRs, administrative claims, registries, hospital charge data sources, patient-generated data and others) are on the rise with thousands of studies published in 2021 alone [84]. Observational data are used in clinical decision support tools including more traditional evidence aggregative tools and more novel applications of natural language processing and machine learning in risk-prediction models and early warning systems [85]. Among other uses of observational data, observational effectiveness and safety studies have been shown to inform clinical decision-making by clinicians and regulatory bodies [38,56,83,86,87].

Despite its common use, observational research is criticized for potential residual confounding [67]. Validity can be compromised at each step of observational study (Figure 2) potentially leading to faulty evidence.

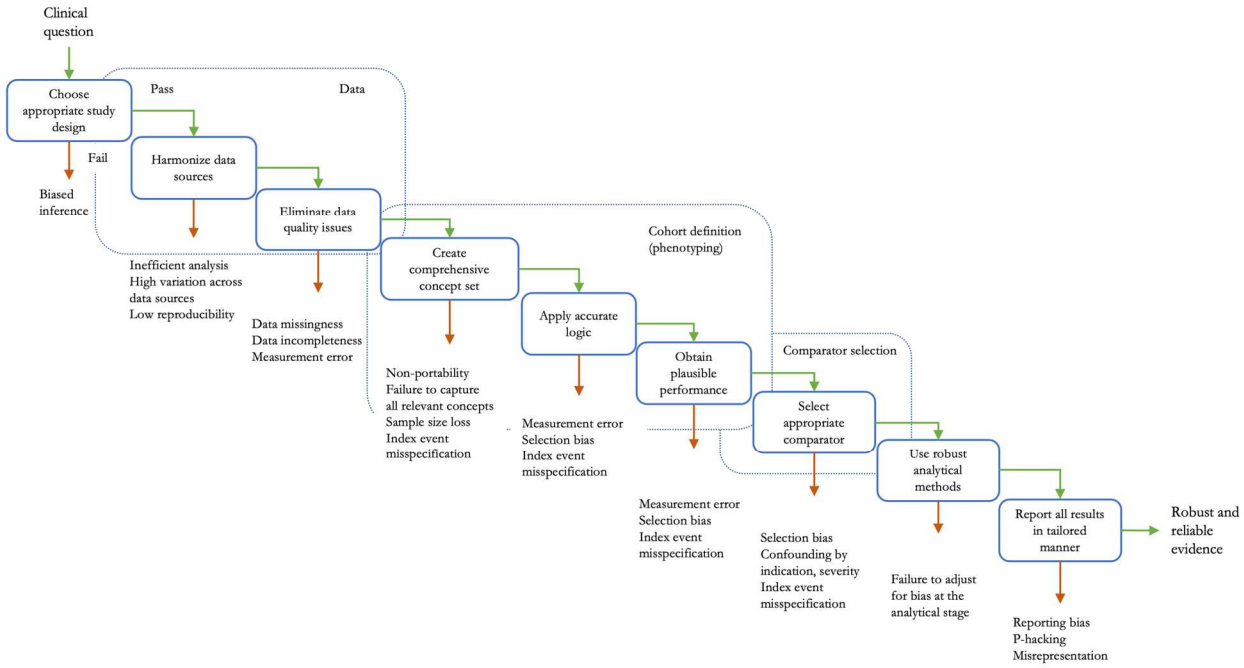


Figure 2. Threats to validity in observational safety and effectiveness studies.

Appropriately addressing these threats is required but substantially lengthens evidence generation. While the promise of observational research is that it can be completed at a fraction of RCTs' cost and time, studies take years from conception to publication [72,88].

We will briefly discuss threats to validity with a more in-depth discussion of the challenges and informatics solutions in the next section of the thesis as well as in the individual subsections.

Study design

Selecting an appropriate study design plays an important role in study validity. Generally, formulating research question according to the Population, Intervention, Comparator, Outcome, and Timing (PICOT) framework [89] helps to identify the appropriate design and map a clinical question to a set of conceptual elements to be later transformed into an executable algorithm. While seemingly straightforward, this process requires understanding of the potential biases associated with a given study

design (be it a cohort study, a case-control study, a self-controlled case series or else) and relevance and applicability of the data source at one's disposal.

The recommendations for selecting the appropriate study design [64,90] and empirical studies of different study designs [91] can be found elsewhere. In this work we mostly focus on cohort design as the most commonly used and less bias-prone in most of the scenarios [91–93].

Data

As this thesis concerns retrospective secondary analysis of existing data captured for clinical or billing purposes, the design of an observational study and its validity should be directly based on the ability of the data to support a given question.

As opposed to the tightly controlled data collection in prospective studies, observational data are sparse and clinical observations are oftentimes missing not at random [94]. The data are generated only when a patient interacts with the healthcare system and patients are observed more frequently when they are sick, so that the timing of care and patient evaluation is highly irregular compared to the standardized protocols in prospective studies.

EHR data (except for national healthcare systems) reflects fragmented care where care for certain disorders or emergency conditions is sought outside of the institution and therefore is not captured [88].

Completeness and accuracy of data capture is limited and varies greatly across different data sources with some common patterns driven by the data provenance [95]. For example, rich EHR data more commonly contains information about the results of clinical measurements and tests, patient-reported data, and socio-economic variables. On the other hand, administrative claims data can provide a more comprehensive data capture since it follows the patient at different institutions. It should be noted

that despite the common patterns, there is a substantial heterogeneity in the data sources within the same country, provenance, or level of care.

Depending on the data source, some of the elements such as outpatient prescriptions, inpatient visits, socio-economic variables or diseases with privacy concerns can be missing, which can make the data source irrelevant for a given research question. Broadly speaking, relevance of the data (including characteristics of the population, average follow-up and available data elements) needs to be assessed in every study and oftentimes requires extensive data exploration [57]. Researchers and regulatory bodies argue that any study should be accompanied by in-depth analysis of the relevance of the data source, including a description of coding practices [57], which makes any feasibility study time-consuming. As of this moment, there is no transparent and efficient way to determine if a given data source is suitable for a particular question or quantify the impact of local practices on patient selection and study estimates.

Aside from low relevancy, validity of observation studies can be undermined by low data quality, which can be measured across three axes: conformance, completeness and plausibility [96]. Briefly, as opposed to clinical trial data, high volume of EHR and claims data is oftentimes associated with manual entry errors or artifacts of care delivery process such as copy and paste information propagated from old records [97].

Common examples include erroneous entries, especially for poorly standardized fields like laboratory test results, erroneous code assignment, implausible values in measurements and drug prescriptions, missing dose or quantity for drug administrations. On top of that, other specifics of data collection need to be considered when designing a study. For example, as opposed to administrative claims where the enrollment period or coverage is explicitly defined, use of EHR data requires inferring

observation period [63]. The accuracy of an observation period definition is especially relevant for defining incident cases, capturing the events during follow-up, or estimating loss to follow-up.

There are numerous research works and tools for improving data quality directly or indirectly, including standards and frameworks, interface terminologies, methods for inferring missing data and informatics tools for quality checks [96,98–103]. Yet, the challenge of systematic assessment of reliability and relevance of the data has not been fully solved [104]. ***Section 3.4*** explores some of the data quality issues we encountered.

Measurement error

Limited completeness and accuracy of observational data influences our ability to accurately capture exposures, outcomes, and covariates, oftentimes introducing measurement error (also called information bias and misclassification).

Researchers identify patients and events of interest with computable phenotyping algorithms that aim at transforming the raw data into clinically relevant features [84]. Phenotyping algorithms are operational executable definitions constructed based on the conceptual definition of a state, condition, or exposure reflecting the scientific and medical knowledge [57]. Such algorithms are developed based on the available data (structured or unstructured) and depend on the quality of data as well as on local care policies, coding and clinical practices [105].

Measurement error can distort the observed relationship between an exposure and outcome in any direction or affect the precision of the estimates [106] and can be introduced at any stage of algorithm development.

First, it can be introduced at the stage of selecting concepts that represent the clinical idea of interest. Such concepts can come from billing data (such as ICD-10(CM) or CPT-4 codes) or from other sources like unstructured notes mapped to concepts in ontologies (such as SNOMED-CT or RxNorm). Given variability of coding practices, heterogeneity of data in different institutions and large ontology space, ensuring that all relevant concepts are captured is not a trivial task [107]. Concepts and concept sets (code sets, code lists, value sets) are commonly borrowed from the literature with an assumption that they will capture the patients of interest in one's data source. Still, this assumption oftentimes does not hold even within the institutions in the same country providing similar level and scope of care [108]. If concept sets are developed de-novo, extensive exploration and validation is usually required and is ideally performed by a team of data and clinical experts [57].

Promising developments in semi-supervised and unsupervised concept selection allow to achieve better speed and objectivity on the local data but the developed concept sets may not be readily transferable to other institutions [109]. As a result, the process of creating code sets is highly variable and rarely robust. As validation is typically performed on the phenotype algorithm level, the influence of code selection on patient characteristics is rarely studied and the extent to which it introduces bias in the study remains largely unmeasured. We provide more details on the challenges in ***Section 3.1 and 3.2.***

Second, measurement error can be introduced when the conceptual definition is translated into an operational definition. In rule-based phenotyping, which remains the most common approach and whose interpretability adds trustworthiness, an operational definition represents a set of Boolean or temporal rules applied to the concept sets and, similarly to the inclusion and exclusion criteria in RCTs, influences patient selection. Observational studies using the same data and analytics may show different results depending on how the conceptual definition was operationalized [110].

While there is an agreement that this process must be transparent and reproducible, the current research on generalizability and portability of algorithms is not comprehensive. We can note a growing body of standards and recommendations (partially supported by informatics tools) for reproducible phenotype algorithms, its storage and reuse [111–114]. On the other hand, studies of portability of both expert-based and data-driven phenotypes (discussed in more detail in the subsequent sections) remain sparse.

Measurement error is commonly quantified through estimating performance of a phenotype against a gold standard (chart review or another external source of truth). Some argue that the most rigorous approach is to verify all study variables for each patient, including exposure, outcome and all covariates [57], which is almost impossible even on small data sets. If performed, validation is commonly done only for outcomes and positive predictive value (but not sensitivity or specificity) is calculated [115].

Manual chart review is time- and labor-intensive, variable and subjective. Numerous research works attempt to establish probabilistic approaches to phenotype evaluation, compensate for a typical small sample size or develop tools for more efficient information retrieval. Still, the status quo remains unchanged.

Aside from major limitations of current gold standard validation, the research on incorporation of measurement error into study estimates has not been widely adopted in effectiveness and safety studies [106,116]. Current recommendations state that the impact of misclassification and error on study validity should be scrutinized to assess its degree and implications, but most papers simply acknowledge a possibility of measurement error and misclassification, stating that they are not likely to be differential.

Aside from outcome and exposure misclassification, bias can stem from index event misspecification, which represents correct classification of patients, but incorrect assessment of condition or exposure start date. Index event misspecification (discussed in ***Sections 3.2 and 4.1***) can, in turn, influence study estimates, especially in survival analysis. Nevertheless, it is largely underappreciated and is mostly discussed in the context of prevalent versus incident cases [117,118] or disease sub-classes [119].

Bias in comparator definition

Aside from the error coming from incorrect classification of exposure or outcome, there are other biases that must be recognized and mitigated at the stage of developing the study design and specifying target and comparator group or time (Figure 2) as the statistical approaches may fail to address them [120].

They are most often arising from a failure to identify an appropriate comparator (a cohort or time) where the comparator should serve as a proxy for a counterfactual of the target (exposed) population. In RCTs, the treatment is given at random. In real world, as reflected in observational data, administration of a drug depends on multiple factors oftentimes not captured in the data: physician or patient preference, compliance, ability to pay, or expected survival, so that patients receiving different treatments can be fundamentally different.

As ***Chapter 4*** addresses pre-analysis bias in comparator definitions, we will briefly outline the examples of the biases. Outcome detection bias or outcome misclassification bias may occur if the outcome is ascertained differently in study groups, for example due to an awareness about potential adverse events [66]. As opposed to the measurement error we discussed previously, this bias (as well as those discussed below) exists in the context of comparator selection [91].

Confounding by indication, severity or frailty is common and hard to control for as the information about indications, disease severity and patient state is oftentimes not recorded in the structured data. Broadly, it is a distortion of the exposure-outcome association stemming from the difference in the distribution of the variables that influence the outcome in the study groups [64]. As a result, the probability of receiving one of the treatments is not independent from the probability of developing the outcome. For example, confounding by indication or severity can be observed if one of the compared drugs is prescribed in advanced disease while the other is prescribed to the patients with mild symptoms and relatively uncomplicated course of the disease [121]. Similarly, confounding by contraindication can be observed if a particular drug is known to cause a specific complication or side effect that happens to be the outcome in the study [122]. Health-seeking behavior-related biases are also common with certain medications such as preventative treatments being associated with a different health seeking behavior in their users [123]. A classic example is healthy user or vaccinee effect, where patients receiving preventative treatment have better outcomes as they are more likely to partake in other healthy behaviors that are oftentimes not captured at the data [124].

On the other hand, confounding by frailty can occur if preventative (or other types of) treatment is not given to those patients who are likely to die before they benefit from it [125].

A large group of time-related biases include immortal time bias, misclassification of the exposure effect window or depletion of the susceptible [65,126] or time-varying confounding, which occurs if there are temporal patterns in how confounders are recorded. Depletion of the susceptible is common in vaccine effectiveness research (which we will touch on in ***Section 4.2***) if the patients at higher risk of the outcome are depleted from the at-risk population at different rates in study groups [127].

Immortal time bias is common when the time-at-risk start is not aligned between the target and comparator groups, for example when one of the groups requires additional exposure, which, by design, forces the inclusion of patients with longer follow-up time available [128]. It can be a particular challenge in the studies that compare exposed patients to unexposed (users to non-users).

Having an unexposed (non-active) comparator or unexposed time is especially bias-prone [90]. Among the several types of controls (active comparator, inactive comparator, non-user or unexposed comparator, historical control) the preference is usually given to the active comparator as comparing the drugs within the same class or with similar indication should minimize confounding [129,130]. A comparator group should reflect clinically meaningful choices [131], and it is not always possible to choose an active comparator, especially for preventative treatments [130]. If no appropriate comparator exists, some drug safety surveillance studies explore the use of the drugs with different indications as a proxy for unexposed comparator or placebo [132]. Some authors argue that observational studies with unexposed comparators have low validity [88], as with unexposed comparators the selection strategy and index date (also referred to as time zero) are not clearly defined and are left up to a researcher's discretion [130]. Careful consideration is recommended when setting the index date and criteria for the unexposed comparator, but the details of this process and its exact operationalization remains largely unclear [90,130].

Over the years, there has been a number of analytical approaches aimed at mitigating observed bias, from different flavors of adjustment in outcome models to different flavors of matching and stratification, including propensity score and disease score models (further discussed in the next section) [133–137]. Nevertheless, there is still a concern that unmeasured confounding may be present in any study undermining reliability of the evidence [138]. More informatics solutions are needed for

comparator (especially unexposed comparator) definition to attempt to systematically assess, report and mitigate pre-analysis bias.

Other biases and challenges

Besides pre-analysis bias, other challenges undermining reliability of observational research (corresponding to the last box in Figure 2) include p-hacking, publication bias and non-reproducibility. P-hacking and publication bias influence the perception of the results of observational studies as only statistically significant results tend to be published [139]. There are many solutions proposed including standardized protocols for study execution and reporting such as STROBE, RECORD, ENCePP, and TRIPOD guidelines [76,140–142], but we still observe a lack of adoption of these solutions [143–145].

All or some of these challenges may be encountered when running observational studies on individual data sources. There are also shared issues that obstruct valid inference: limited scope of care, coverage, and rather small number of patients, so that a lot of important outcomes and exposures, including but not limited to orphan and new drugs and rare diseases cannot be reliably studied in a single data source. Pooling the data from multiple data sources can meet the sample-size requirement with distributed observational networks effectively collecting the aggregated study results from individual data sources while maintaining patient privacy. Networks enable effective collaboration, which results in the development of novel tools and methods that are used both within the networks and in the separate institutions.

Nevertheless, such distributed analyses pose new challenges to valid inference. In the next section we will discuss the existing distributed observational data networks and their approaches to mitigating bias.

2.2 Distributed observational data networks: challenges and advances

Distributed data networks (federated networks, big-data networks), in the US and overseas, partially solve the problems of individual data sources. In networks, the data are stored and maintained by each individual data partner, which offers benefits in terms of potential number of data sources involved and patient privacy compliance [146].

Current networks begin their history in the early 2000s, when multiple research institutions gathered together to address questions of drug surveillance (OHDSI, EU-ADR, Vaccine Safety Datalink and Sentinel) [147–150], genomics research (eMERGE) [151], clinical research in children (PEDSnet and Capricorn) [152,153], use of biological and clinical data (Informatics to bedside, I2B2) [154], diversity and inclusion [155], COVID-19 research [156,157] and pragmatic clinical trials [158]. While these were the initial use cases, networks support studies of rare diseases and outcomes, new drugs and diverse patient population, incorporate different modalities of data (structured, unstructured, genomics, clinical trials data), develop new methods and more. They are currently used for adverse event surveillance and post-marketing monitoring, clinical trial enrollment, monitoring of adoption of practices and drug utilization studies. Network studies have shown a potential to support and reinforce the results of the RCTs as well as complement them [159].

The networks enable performing multiple comparative effectiveness studies at once, potentially making the evidence generation process fast and efficient while maintaining reliability and generalizability of findings [45,169]. On the other hand, distributed data networks share common challenges, which include data harmonization across the institutions, addressing data heterogeneity, development of scalable and reusable phenotyping algorithms and analytical pipelines, along with a broad challenge of creating infrastructure for data transformation, access, study execution and reporting.

Leveraging diverse and large populations in large distributed observational networks reduces random but propagates systematic error [160], which motivates the research on scalable, re-usable and portable methods for measured and unmeasured confounding.

The remainder of this section as well as the thesis overall specifically focuses on these challenges and their influence on timely and reliable evidence generation.

Data standardization and harmonization

The first networks' contribution to scalable and reliable research is common data models (CDMs). All networks establish common data models with a varying degree of fidelity to enable portable research across all participating data sources [161,162]. Common data models specify the structure and format of the data so that unified approaches to data access, processing and analysis can be used. Some (Sentinel, Clinical Data Interchange Standards Consortium Study Data Tabulation Model [CDISC SDTM], PCORnet) provide syntactic mapping while preserving local data collection and processing rules and ontologies, and others (OHDSI) also perform semantic standardization of the content by providing a common reference ontology system [163,164].

With syntactic mapping, standardization happens at the analytical stage with the source data left unchanged. With semantic standardization, data transformation happens at the extract-transform-load (ETL) stage with all data sources complying to a common data and ontology standards. The latter is especially important in international research, where participating data sources established their own coding systems for diagnoses, procedures and medication [165].

Aside from providing means for unified analytical approaches, CDMs enable scalable data quality assurance procedures to ensure internal validity. Within each network, there are programs, working groups and initiatives that regulate data quality: Sentinel's Data Quality Review and

Characterization Programs [98], OHDSI's Themis and Data Quality Dashboard [101], HMORN's VDW QA checks

[99], etc. CDM allows for standardized and scalable quality assurance procedures that focus on data completeness, integrity, conformance and plausibility at scale rather than developing standards for individual data sources or unit cases for separate conditions [101]. Some of the examples of procedures that are crucial for study validity include checks for missingness, evaluation of summary statistics for laboratory test results, evaluating the frequency and trends in diagnosis and procedure occurrences.

The field of the data quality in networks is constantly evolving with new frameworks and methods being developed [96,102]. Since the quality checks are usually mandatory for all data partners, they set up a high bar for data quality, which oftentimes cannot be achieved in individual data sources and uncover data challenges that are easily overlooked by the local data experts.

Identification of patients of interest (phenotyping)

In this overview, we do not cover a broader research literature on phenotyping. Instead, we focus on specific research in networks that addresses the challenges of developing and validating scalable and portable phenotypes across disparate institutions.

Historically phenotyping in networks has been rule-based, where the choice of the elements and logic was driven by the experts [166–168]. Extensive collaboration of clinicians and informaticians within multiple institutions greatly contributed to assessing measurement error and bias in phenotyping algorithms. Researchers highlighted the challenges related to using disparate data sources and a need for iterative development [167,169]. This led to new standards for executable phenotypes such as desiderata for phenotype algorithms, which focus on the standardization of representations, use of common terminologies and data models [2] as well as infrastructures that supports phenotype authoring,

documentation, execution, validation and re-use (PhenoFlow and PheMA, CALIBER data portal, VAPheLib, PheP and Phenotype Library) [111–113,170–173].

Selection of concept sets in networks is challenging and time-consuming given the disparities in patient representations at the different institutions. One common approach has been to leverage local knowledge and develop concept sets that are source-specific to reflect local concept use and billing practices. Such an approach is especially common for organizing CDMs and provides high level of customization at a cost of low consistency and scalability [162,166,167].

Other advances in rule-based and mixed-method concept selection include empirical exploration of the local data to learn and extrapolate the patterns of code usage, which improves robustness of code selection by leveraging the data to guide code selection, but still lacks scalability [167,169].

In pure data-driven code selection (or, as commonly called, feature selection) there are several advances made in using multi-view banded spectral clustering [174], non-negative tensor-factorization [175], latent Dirichlet allocation [176] or leveraging knowledge graph-enhanced embeddings [177] to guide code selection. While structured data (such as ICD-10 codes) remains the most common modality of the data used, the researchers also leverage notes and external knowledge sources with surrogate-assisted features [178], or notes with sparse embeddings [179] to select appropriate concepts to represent a clinical idea.

As we discussed before, the concept sets are used as building blocks used in conjunction with Boolean or temporal rules to create an operational cohort definition. The body of research focusing on sensitivity of patient selection to the choice of such rules and the order of their application is rather limited. As highlighted before, the process is highly iterative with patient characteristics or phenotype performance (positive predictive value, sensitivity and specificity) being the criteria for selecting the

optimal phenotyping algorithm [73,169]. Nevertheless, the design choices remain poorly described in most of the published literature and the influence of such choices on patient selection and study estimates remains largely unexplored. This is especially relevant for the studies that involve international data sources, as the research has historically been US-centric.

Structured data can be supplemented by the other types of data such as FAERS for adverse events [180], unstructured notes and reports [181–184], or patient-reported data [185]. While promising, the latter approaches require appropriate infrastructure and availability of data elements and may not be generalizable to the whole network. Since the natural language processing (NLP) or machine learning systems need to be portable across the institutions, several groups focus on developing flexible NLP models and frameworks to be executed against the network [184,186,187] as well as more complex super-learner frameworks combining the information from the notes and external knowledge sources [188] and tools enabling XPRESS framework with dynamic anchor learning [189,190].

These research works specifically account for the challenges associated with developing portable phenotypes that can be executed and show similar performance across disparate disjoint institutions.

As noted before, learned knowledge is biased to specific institution [109], making phenotype portability a key challenge in large-scale observational research.

Based on our assessment, portability may be influenced by three broad groups of factors: (a) population (age, employment, race, ethnicity, prevalence of disorders), (b) time period, which impacts vocabularies used in the data and prevalence of disorders and (c) data source specifics (billing practices, level of missingness, scope of care and others).

While it has been extensively studied for individual conditions such as asthma, chronic kidney disorder or diabetes [153,163,191–194], there is a dearth of systematic research on factors influencing

portability and potential biases associated with a lack of thereof. Previous studies concerned the impact of data fragmentation [169,195], data provenance [196–198], ontologies and semantic homogeneity [194,195] and studied populations [199,200] on portability of both expert-based and data-driven phenotypes. Overall, they concluded that portability may be subject to reporting bias and does not guarantee future portability, especially among international data sources.

Overall, the studies conducted in the networks merely acknowledge the fact that misclassification, selection, or other bias may exist when choosing a comparator and address them by adding sensitivity analyses that test the same hypothesis but have different study design parameters [201–203]. Given that the bias can propagate when multiple disparate data sources are used, there is a need for systematic assessment of different strategies for defining unexposed comparator and the influence of such strategies on patient characteristics and study estimates.

Analytical methods and tools

In reliable evidence generation, the next step after taking appropriate measures to mitigate bias at the stages of design, patient identification and comparator selection is to apply robust analytical methods to address residual bias.

First, given the non-randomized nature of observational studies, observational studies in networks commonly use variations of propensity score modelling to adjust for measured confounding.

The subjects in the treatment and comparator groups are stratified or matched on their propensity score (probability of a subject receiving one treatment instead of the other conditional on baseline characteristics), which allows to adjust for confounding in a similar fashion as done in randomized experiments [204].

While in the studies conducted on individual data sources researchers oftentimes select several covariates for propensity score model based on their knowledge or previous literature, such an approach oftentimes fails to adjust for confounding in large-scale studies with thousands of observed covariates. Some networks adopted methods for empirical selection of relevant covariates such as high-dimensional propensity score adjustment [205,206]. Others use large-scale propensity score matching (LSPS), which also uses empirical selection but selects a large number of covariates (tens of thousands) and is superior to selecting a small set of hand-picked covariates [207–209].

LSPS has been shown to adjust for missing variables if such variables may correlate with the observed ones [20]. This can be especially important for addressing the concerns about the validity of observational evidence given the data missingness and a common lack of socio-economic variables. Other alternatives include probability of treatment weights [210], instrumental variable analysis [211] and cardinality matching that showed superior performance to LSPS for rare conditions or orphan drugs [212].

The networks use approaches to quantify residual study bias due to unmeasured confounding and incorporate it into study estimates. For example, negative controls (well-studied controls for which no known relationship to the outcome or exposure exists), negative control risk periods (time windows in which the exposure has no biological effect) and synthetic positive controls are used to assess residual bias and calibrate p-values and confidence intervals of the estimates [160,204,210,213].

Second, evidence diagnostics play a critical role in making sure that the evidence is reliable and ready to be supplied to clinicians. The networks contributed greatly to the development of standardized pipelines and tools for assessing the quality of evidence, including assessing covariate balance distribution, preference score distribution, assessment of the estimates from negative controls and more.

Preference score plots are used to assess the degree to which the treatment and comparator groups have similar baseline characteristics [214]. Poor equipoise in a given treatment-comparator pair indicates that only a small proportion of patients in the groups have similar baseline characteristics and, therefore, only a small portion of patients are eligible to be included in a valid comparison.

Covariate balance plots are used to inspect the covariates that do not meet the threshold for standardized mean difference and are said to be unbalanced, reflecting potential confounding [132,215]. Other tools include Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD), which can be used to detect protopathic bias or misclassification of the dates of the adverse events and eliminate false-positive drug-event associations. [216,217]

Finally, as the networks enable generation of multiple estimates for a given hypothesis, a body of research focuses on accurate assessment of heterogeneity of estimates, which can point at a data source-specific bias, as well as on methods for producing a single estimate of effect from multiple institutions. Along with traditional meta-analysis approaches, the networks also propose the use of distributed models such as distributed multivariate regression [218].

Gaps addressed in this thesis

Observational studies are prone to measurement error and bias with analytical approaches only partially addressing them. Potential bias can lead to unreliability of evidence, lack of credibility and addressing it currently is not standardized and time-consuming. Current research addresses the questions of reproducibility, replicability, transparency, and robustness of real-world evidence, but the problems of generalizability and scalability remaining largely unsolved. These problems are especially prominent in large distributed observational networks that gather data from disparate institutions and populations capturing different aspects of care. There is a solid body of research on different sources of bias and its

implication, but there is the magnitude of bias in networks remains underexplored. Accordingly, readily available informatics solutions to systematically assess bias at the pre-analysis (design) stage do not exist.

First, efficient and scalable phenotyping remains the bottleneck in the studies conducted in large observational networks. Phenotype development remains highly iterative and time-consuming with an average phenotype taking months to develop. Given that efficient pipelines for phenotype development and evaluation have not been adopted, the current research, both expert-based and data-driven is limited to separate conditions and use cases.

In networks, the approaches to concept set creation and cohort definition remain variable and poorly described, so the influence of design and operationalization choices on patient selection and study validity is not fully explored. Data-driven approaches have a potential to be more scalable, but their portability, efficiency, and potential to influence decision-making are yet to be shown.

While portability and generalizability are critical in efficient phenotyping in networks, there is lack of studies on factors influencing portability and, moreover, lack of informatics solutions to ensure such. The practice of borrowing phenotype definitions and concept sets from published studies remains common potentially jeopardizing reliability of evidence. On the other hand, data-driven phenotyping largely remains limited to one data source with no available solutions to ensure generalizability to other data sources in a network.

Similarly, there are no readily available solutions to efficient phenotype evaluation in distributed networks. If performed, it is done through manual review of a small sample of charts, which is neither reliable nor scalable. Data-driven approaches, on the other hand, provide scalability and ability to evaluate phenotypes more comprehensively but lack interpretability and broad adoption.

Second, there is a lack of systematic approaches to comparator definition, especially when the comparator represents unexposed patients. Comparator definition choices may compromise validity of evidence as any deviation from the population of interest (such as in demographic characteristics) can introduce bias. Nevertheless, impact of comparator definition on patient selection and study estimates is rarely studied and remains mostly unknown in the settings of heterogeneous data sources in networks. Choice of comparator(s), especially unexposed, is mostly guided by the expert knowledge rather than data and it is unknown to what extent the bias introduced by such a choice can be controlled for by analytical methods.

Establishing standardized and systematic pipelines for phenotype development, evaluation and comparator definition would enable scalable, reliable, and more efficient research in observational networks and, potentially, would reduce time and cost of evidence generation.

2.3 Introduction to OHDSI and data sources used

Throughout this work we take advantage of, build upon, and contribute to the methods and practices in Observational Health Data Sciences and Informatics (OHDSI) – international multi-stakeholder, interdisciplinary data network of electronic health records, administrative claims, hospital discharge data, registries, and other observational data sources. Given its importance in understanding the content of this thesis, we will briefly describe its principles and summarize the data sources within the OHDSI network that were used in this thesis.

OHDSI encompasses more than 900 million unique patient records across 40 countries and is a semantically harmonized network [147]. OHDSI's Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) stores the data from different sources (electronic health records, claims data, registries, surveys, trial data, etc.) and geographies (such as US, Asia, Europe, Australia) in a unified format guided by the ETL (extract-transform-load) procedures and policies. It provides both syntactic and semantic standardization through the CDM that organizes the data into a common relational database structure and the OHDSI Standardized Vocabularies that provide a common reference ontology system to harmonize the data content.

As we discussed, the common data model supports distributed research at scale. While it comes at a cost of spending resources on converting the data, the conversion itself has been shown to have little impact on study results [163,164]. Currently, there are multiple published papers on peculiarities of transforming raw data to the OMOP CDM [219–232], but the principles behind the CDM are not unique to OHDSI.

On the other hand, OHDSI Standardized Vocabularies, which is a backbone of OMOP CDM, are unique to OHDSI. They specify a priori representation for patient data and serve as a reference the OMOP CDM so all data are represented in a structured format avoiding free text fields. As opposed to the other terminology systems designed around information retrieval [233,234], the goal of the Standardized Vocabularies is to provide (a) a standard for representing the data content and (b) support for the mapping of various vocabularies and coding schemes adopted in the data sources within the network.

OHDSI Standardized Vocabularies are created de-novo or imported from existing US and non-US taxonomies, terminologies, ontologies, and vocabularies and cover each main medical domain: diagnosis ("Condition"); laboratory and instrumental tests ("Measurement"); medical procedures ("Procedure"), medications ("Drug"), medical devices and supplies ("Device"); and clinical signs, symptoms and observations ("Observation"). Patient-level data in OMOP CDM is coded using the concepts from a subset of standard vocabularies inside the OHDSI Standardized Vocabularies with other concepts being linked ("mapped" to the standard ones) [101]. Having standard reference terminologies across all data sources enables standardized approaches to phenotyping, feature extraction and large-scale analytics.

Specifically, semantic standardization in the OHDSI Standardized Vocabularies is achieved by defining all concepts through their relationships and selecting one referent concept per semantic meaning. For example, the majority of codes from ICD-10(CM), ICD-9(CM), their international flavors, Read and ICD-O-3 fall under Condition domain and are mapped to referent concepts in SNOMED-CT (Figure 3).

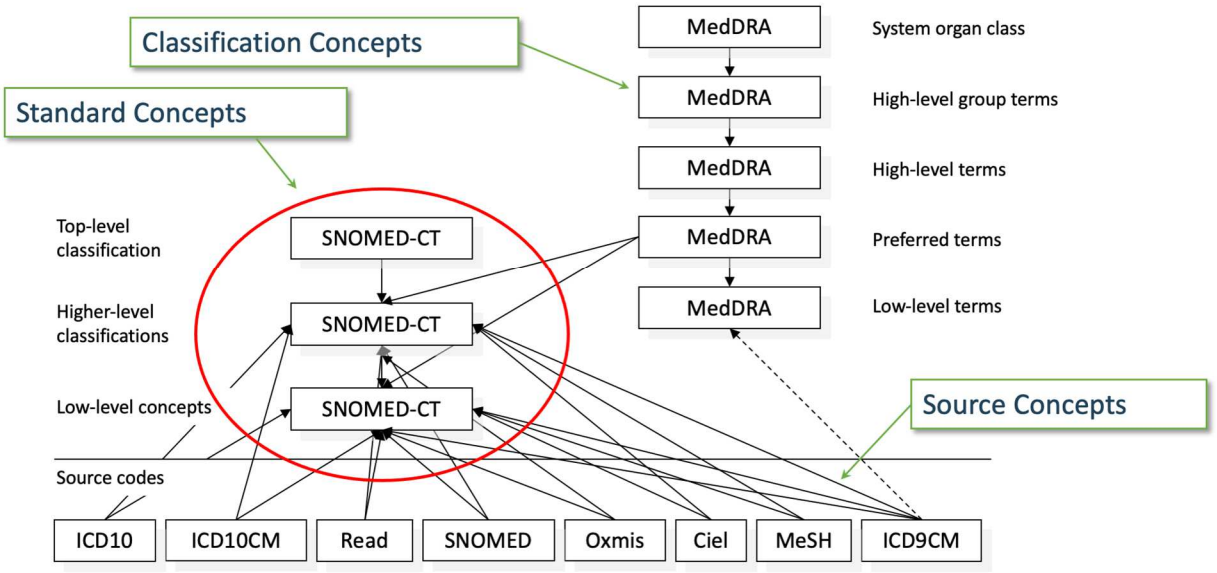


Figure 3. Selected elements of the OHDSI Standardized Vocabularies in the Condition domain.

Throughout this thesis, we leverage existing OHDSI Standardized Vocabularies structures, such as joint drug hierarchy of imported RxNorm, created de-novo RxNorm Extension, Anatomical Therapeutic Chemical (ATC) classification and CVX; partially aligned procedure hierarchy of SNOMED-CT, CPT-4 and HCPCS and other elements. While not covered in detail, we contributed to these endeavors [32,235–237], which both enabled us to fully appreciate their utility in observational research and highlighted complexity of medical ontologies, which limits their full use by a broader community. While it has been shown that mapping to standard terminologies in OMOP does not introduce significant error [164], proper use of OMOP Standard Vocabularies still requires extensive knowledge.

Throughout this thesis, we use data sources converted to OMOP CDM version 5. Table 1 outlines the data sources used in each section. For almost each section, the data sources were engaged

separately by designing a study, writing a study protocol, creating study packages and advertising projects on the OHDSI forum or person-to-person.

Table 1. Description of the data sources used in this thesis. EHR – electronic health record data, claims – administrative claims data.

Name	Type	Country	Size	Description	Where used
Ajou University Database	EHR	Korea	3 million	Korean tertiary teaching hospital electronic health record data including visits data, medication administration and prescription information, procedures and diagnosis.	Section 3.1, 3.2
Australian Electronic Practice-based Research Network (AU-ePBRN)	EHR	Australia	0.2 million	Electronic health records data from primary care practices in Australia	Section 3.1, 3.2

Clinical Practice Research Datalink (CPRD)	EHR	UK	12 million	CPRD is a governmental, not-for-profit research service, jointly funded by the NHS National Institute for Health Research and the Medicines and Healthcare products Regulatory Agency. CPRD consists of data collected from UK primary care for all ages. This includes conditions, observations, measurements, and procedures that the general practitioner is made aware of in additional to any prescriptions as prescribed by the general practitioner. In addition to primary care, there are also linked secondary care records for a small number of people. The major data elements contained within this database are outpatient prescriptions given by the general practitioner and outpatient clinical, referral, immunization or test events that the general practitioner knows about.	Section 4.1
Columbia University Irving Medical Center (CUIMC)	EHR	USA	6 million	The Columbia University Irving Medical Center (CUIMC) database comprises electronic health records on more than 6 million patients, with data collection starting in 1985. CUIMC is a Northeast US quaternary care center with primary care practices in northern Manhattan and	Chapters 3-5

				surrounding areas, and the database includes inpatient and outpatient care. The database currently holds information about the person (demographics), visits (inpatient and outpatient), conditions (billing diagnoses and problem lists), drugs (outpatient prescriptions and inpatient orders and administrations), devices, measurements (laboratory tests and vital signs), and other observations (symptoms). The data sources include current and previous electronic health record systems (homegrown Clinical Information System, homegrown WebCIS, Allscripts Sunrise Clinical Manager, Allscripts TouchWorks, Epic Systems), administrative systems (IBM PCS-ADS, Eagle Registration, IDX Systems, Epic Systems), and ancillary systems (homegrown LIS, Sunquest, Cerner Laboratory).	
IBM MarketScan	Claims	USA	144 million	IBM MarketScan Commercial Claims and Encounters Database (CCAE) is a US employer-based private-payer administrative claims database. The data include adjudicated health insurance claims (e.g., inpatient,	Section 4.1, 5.2

Commercial Claims and Encounters Database (CCAE)				outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. Additionally, it captures laboratory tests for a subset of the covered lives. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.	
IBM MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR)	Claims	USA	10 million	IBM MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR) represents health services of retirees in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. These data include adjudicated health insurance claims (e.g., inpatient, outpatient, and outpatient pharmacy).Additionally, it captures laboratory tests for a subset of the covered lives.	Section 4.1, 5.2

IBM MarketScan Multi-State Medicaid Database (MDCD)	Claims	USA	28 million	IBM MarketScan Multi-State Medicaid Database (MDCD) contains adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. Members maintain their same identifier even if they leave the system for a brief period; however, the dataset lacks laboratory data.	Section 4.1, 5.2
Information System for Research in Primary Care – Hospitalization Linked Data (SIDIAP-H)	EHR	Spain	2 million	Primary care records database from Catalonia, North-East Spain. The SIDIAP-H subset of the database includes around 2 million people out of the total 7 million in SIDIAP that are registered in primary care practices with linked hospital inpatient data available as obtained from the Catalan Institute of Health hospitals. Healthcare is universal and tax-payer funded in the region, and primary care physicians are gatekeepers for all care and responsible for repeat prescriptions.	Section 4.1

IQVIA Ambulatory Electronic Medical Record (AmbEMR)	EHR	USA	42 million	US ambulatory records that capture outpatient visits with approximately 50% primary care, 50% specialists. The database contains deidentified medical records and encounters from 100,000 physicians and 800 networks in the United States covering the period from January 2006 through May 2019. These data include provider medical specialty; patient variables such as examination date, year of birth, sex, and race and ethnicity; and clinical variables such as diagnoses, procedures, medication prescription records, and patient and family history captured during a patient visit. Contributing practices consist of medium to large physician offices, outpatient clinics, and physician groups.	Section 3.1, 3.2, 3.4
IQVIA Australia Electronic Medical Records (Australia EMR)	EHR	Australia	6 million	Anonymized patient records of more than 6 million patients in Australia collected from Patient Management software used by GPs during an office visit to document patients' clinical records.	Section 3.1, 3.2, 4.1

IQVIA Disease Analyser Germany (DA Germany, IMSG)	EHR	Germany	34 million	IQVIA DA Germany is collected from extracts of patient management software used by GPs and specialists practicing in ambulatory care settings. Data coverage includes more than 34M distinct person records out of at total population of 80M (42.5%) in the country and collected from 2,734 providers. Dates of service include from 1992 through March 2020.	Section 3.1, 3.2, 3.4, 4.1
IQVIA Disease Analyzer France (DA France, IMSF)	EHR	France	4 million	Electronic health records data from French practices (mostly primary care practices). The data are collected from physician practices and medical centers for all ages.	Section 3.1, 3.2
IQVIA Longitudinal Patient Data France (LPD France)	EHR	France	7.8 million	Anonymized patient records of 7.8 million patients in France collected from Patient Management software used by GPs and select specialists during an office visit to document patients' clinical records	Section 4.1

IQVIA LRxDx US Open Claims (Open Claims)	Claims	USA	160 million	Anonymized, pre-adjudicated claims collected from US office-based physicians and specialists	Section 3.1, 3.2
IQVIA Oncology Electronic Medical Record (OncoEMR)	EHR	USA	2.1 million	The US database capturing oncology ambulatory outpatient visits including laboratory data, procedures, diagnosis, and medication prescription data.	Section 3.1, 3.2, 3.4
IQVIA US Hospital Charge Detail Master (CDM)	Hospital charge	USA	88 million	Anonymized hospital charge detail masters (CDM) collected from short-term, acute-care and non-federal hospitals	Section 3.1, 3.2

Japan Medical Data Center (JMDC)	Claims	Japan	5.7 million	Japan Medical Data Center (JMDC) database consists of data from 60 society-managed health insurance plans covering workers aged 18 to 65 and their dependents (children younger than 18years old and elderly people older than 65 years old). JMDC data includes membership status of the insured people and claims data provided by insurers under contract (e.g. patient-level demographic information, inpatient and outpatient data inclusive of diagnosis and procedures, and prescriptions as dispensed claims information).	Section 3.1, 3.2, 4.1
Korea National Health Insurance Service / National Sample Cohort (NHIS/NSC)	Claims	Korea	1 million	National administrative claims database covering the South Korea population spanning 2002 – 2013 integrating the data from more than 366 institutions. It includes the data about prevention, diagnosis, disease, and injury treatment, as well as rehabilitation, births, deaths and health promotion. Currently the NHIS maintains and stores national records for healthcare utilization and prescriptions.	Section 3.1, 3.2

MIMIC III	EHR	USA	0.04 million	Electronic health records data associated with ~60,000 intensive care unit admission at a large tertiary care hospital. Data includes vital signs, medications, laboratory test, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more.	Section 3.1, 3.2
Optum de-identified Electronic Health Record Dataset (PANTHER)	EHR	USA	94.8 million	The Optum PanTher EHRs are derived from 53 integrated delivery networks from diverse geographies in the US, including more than 700 hospitals and 7000 clinics across the US. Clinical and administrative data are obtained from both inpatient and ambulatory EHRs, practice management systems and numerous other internal systems; and are processed, normalized, and standardized across acute inpatient stays and outpatient visits. Data elements include, but are not limited to, patient demographic information, medications prescribed and administered, laboratory results, and coded diagnoses and procedures.	Section 3.1, 3.2

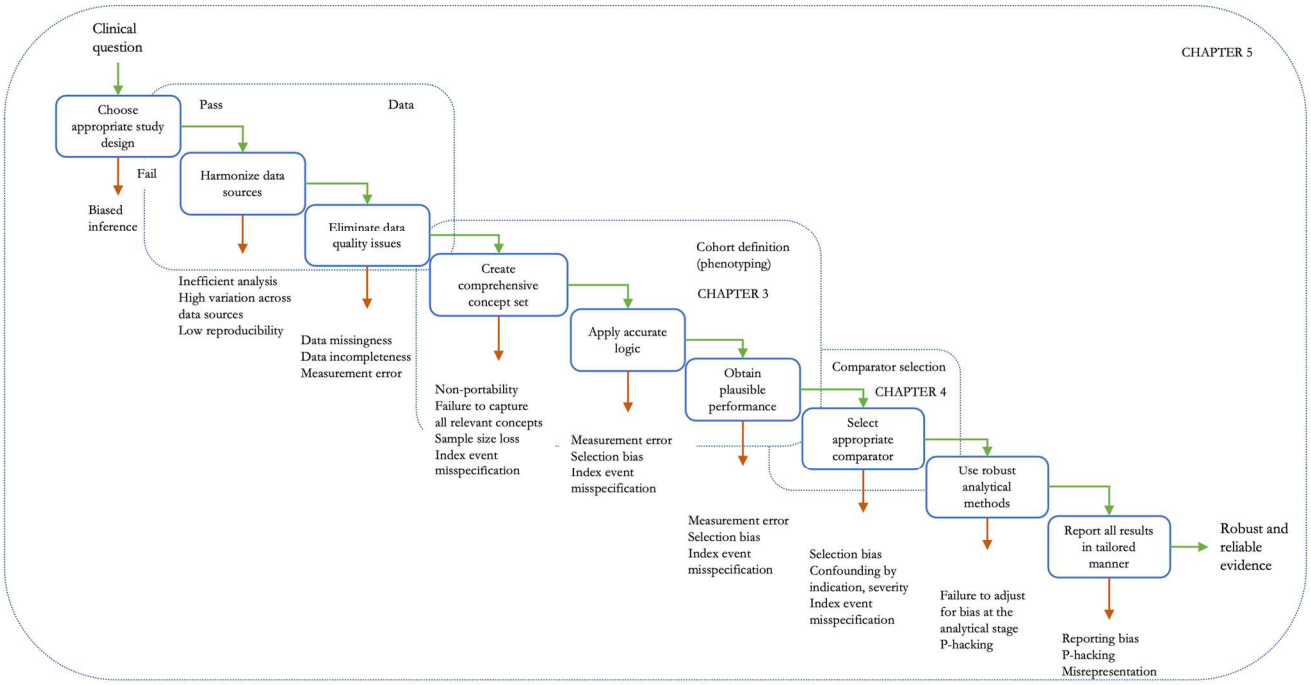
Optum© de-identified Electronic Health Record Dataset (Optum EHR)	EHR	USA	87 million	Optum de-identified Electronic Health Record Dataset is derived from dozens of healthcare provider organizations in the United States (that include more than 700 hospitals and 7,000 clinics). The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical notes using Natural Language Processing.	Section 3.1, 3.2, 4.1
Optum® De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (Optum SES) and Date	Claims	USA	85.8 million	Optum De-Identified Clinformatics Data Mart Database (Optum Insight, Eden Prairie, MN) is an adjudicated administrative health claims database for members with private health insurance, who are fully insured in commercial plans or in administrative services only (ASOs), Legacy Medicare Choice Lives (prior to January 2006), and Medicare Advantage (Medicare Advantage Prescription Drug coverage starting January 2006). The population is primarily representative of US commercial claims patients (0-65 years old) with some Medicare (65+	Section 3.1, 3.2, 4.1

of Death (Optum DOD)				years old) however ages are capped at 90 years. It includes data captured from administrative claims processed from inpatient and outpatient medical services and prescriptions as dispensed, as well as results for outpatient lab tests processed by large national lab vendors who participate in data exchange with Optum. Optum SES provides socio-economic status for members with both medical and pharmacy coverage and location information for patients at the US Census Division level. Optum DOD is primarily representative of US commercial claims patients with full death record.	
Premier Healthcare Database (PHD)	Hospital charge	USA	215 million	Hospital charge data from the hospitals across the US	Section 3.1, 3.2
Stanford Medicine Research Data	EHR	USA	5 million	EHR data derived from outpatient and inpatient visits Stanford Hospital and Clinics	Section 3.1, 3.2

Repository (STaRR)					
The Healthcare Cost and Utilization Project (HCUP), Nationwide Inpatient Sample	Hospital charge	USA	115.9 million	US hospital care data, including inpatient stays, ambulatory surgery and services visits, and emergency department encounters.	Section 3.1, 3.2
The Integrated Primary Care Information (IPCI) database	EHR	Netherlands	2.36 million	Longitudinal observational database containing electronics medical records from a representative sample (n=750) of general practitioners (GPs) in 9 different GP systems in the Netherlands covering up to 10 years of observational data.	Section 4.1
Tufts Medical Center Database	EHR	USA	1 million	EMR data from a large tertiary care hospital, including inpatient, outpatient, state death records, and tumor registry records	Section 4.1, 3.2

Chapter 3. Addressing phenotyping as a source of measurement error

In this chapter we will discuss improving scalability, reliability, and portability of phenotyping algorithms for identifying patients of interest in distributed observational data networks. The figure below contextualizes this step in a retrospective observational study.



While phenotyping on EHR data has been carried out for more than 40 years, the problem of robust, accurate and efficient phenotypes remains largely unsolved. This chapter focuses on assessing and reducing bias in phenotyping in data networks and places a specific emphasis on using data to inform expert-based phenotype development at every step.

First, I investigate code utilization patterns and data source heterogeneity across 22 US and international data sources and describe how it influences phenotype development in networks. I proceed

with developing original methods for computing data source granularity and evaluate them on the network of data sources. I dissect factors affecting granularity and its impact on phenotypes as well as identify the patterns of granularity depending on the provenance of data and country of origin.

Using the dataset generated in the previous step I design and deploy a recommender system for comprehensive code selection across the network. I evaluate its performance for four conditions (diabetes mellitus type I and II, attention deficit hyperactivity disorder and heart failure) on three EHR and claims data sources. I then demonstrate how the recommender system fits into the OHDSI pipeline for phenotype development, evaluation and storage and describe its usage in individual studies as well as 11 major network studies.

Second, I investigate sensitivity of patient composition to the choice of specific algorithmic implementation of inclusion and exclusion criteria (cohort definition). We conduct an experiment with 45 OHDSI collaborators, which was set up as a standardized implementation of the clinical description from a recent high-impact observational study, execute collaborators' implementations and quantify the variability in translation of conceptual definition into an executable algorithm and its impact of on patient selection and characteristics.

Third, we apply the lessons learned in the abovementioned sections to create EHR-derived chronic kidney disorder phenotypes that are portable to administrative claims. We develop four phenotypes with varying complexity and evaluate the algorithms on four US and international data sources and provide recommendations on how to facilitate phenotype portability and validity.

Finally, I identify the gaps in the current approaches for phenotype evaluation. Given the shortcomings of the current gold standard for phenotype evaluation (manual chart review), I present a framework for systematic examination of patient data and apply this framework to patients' structured

data to propose a chart review alternative that enables efficient ascertainment of patient status. I illustrate the utility of the framework for acute and chronic, outpatient and inpatient conditions and demonstrate that, compared to manual chart review, its use achieves better consistency and non-inferior accuracy in classifying patients at a fraction of time.

3.1 Data source heterogeneity, granularity, and code utilization[1]

In data networks, phenotype algorithms are intended to be used across multiple data sources, which requires recognition and reconciliation of differences in patient representations arising from underlying populations, disparate coding practices and specifics of data capture. Nevertheless, there is little research on code utilization across the networks and its influence on patient selection and study validity.

In this section, I collect the data from US and international 22 data sources containing 272 billion records and use this dataset to (a) investigate code utilization patterns, (b) develop original methods for assessing data source granularity and (c) investigate patterns of data source granularity across the data sources with different provenance and country of origin.

I observe high heterogeneity of data sources and discrepancies in coding practices, which plays a crucial role in code selection. I show three SNOMED-based approaches consistently classifying data source granularity and reveal granularity patterns specific to the provenance of data and country of origin.

3.1.1 Background

In observational data networks, studies are executed across data sources with different provenance, country of origin, population, and coding practices. Even when formal semantic interoperability is achieved by standardizing data format (a common data model) and content (standardized vocabularies), there still can be substantial data heterogeneity across sites.

[1] A large portion of this section is published in the AMIA Annual Proceedings 2020. The full citation for this publication is Ostropolets A, Reich C, Ryan P, Weng C et al. Characterizing database granularity using SNOMED-CT hierarchy. *AMIA Annual Proceedings* (2020).

In single-center studies, clinical codes or concepts for phenotype algorithms can be selected based on expert knowledge of local coding practices and data exploration [166,238]. This approach is not suitable for distributed networks as the data from all sites is not readily available. Moreover, the practices in a local institution may not be generalizable to other data sources, especially given that variability in code utilization is not quantified for most of the networks. Current research on impact of coding practices on phenotypes is mostly limited to separate conditions and is limited in scope [105,107]. It is not clear whether these observations can be generalizable to other conditions, nor it is possible to examine code utilization for all conditions separately at scale.

In this section we investigate code utilization and attempt to draw generalizable conclusions about data source granularity, where the latter reflects granularity of the concepts used in a data source.

For example, previously published phenotypes for chronic kidney disease relied on chronic kidney disorder codes (ICD9-CM 585 or ICD10-CM N18 "Chronic kidney disease") [239] or codes with less explicit content, or, in other terms, less granular (ICD9-CM 586, ICD10-CM N18.9 'Renal failure, unspecified' or N19 'Kidney failure') depending on the granularity of the data sources used [240,241].

We hypothesize that the patterns of code use, specifics of population captured, and other factors influence granularity of multiple concepts used in a data source so we can empirically derive the overall level of data source granularity.

3.1.2 Methods

Methods we used for investigating data source heterogeneity and granularity comprised of three elements: (1) data collection across the network, (2) code utilization assessment and (3) data source granularity calculation.

Data collection

To generate the dataset (used here and in the subsequent chapters) we created a call on the OHDSI forum to ask the data partners within the OHDSI network to contribute their data converted to OMOP CDM [147]. Within each site we collected standard and source concepts along with their frequency of use in the data from the main OMOP CDM version 5 tables (Condition Occurrence, Procedure Occurrence, Drug Exposure, Device Exposure, Measurement, Observation) and aggregated them so that the resulting dataset contained the code, aggregated frequency and number of data sources for each concept code.

Hereon, we use the term 'concept', 'code' or 'concept code' to refer to terms in an ontology and 'frequency' to refer to the number of records in a data source.

Data Analysis

Code heterogeneity and utilization

To study code utilization we analyzed the distribution of unique and overlapping concepts in the main OHDSI Standardized Vocabularies domains (Condition, Measurement, Observation, Procedure, Drug), which guide population of OMOP CDM tables and correspond to broad clinical domains: Condition represents all diseases, symptoms and states, Measurement – laboratory tests, vitals and other measurements, Procedure – diagnostic and treatment procedure, Drug – all medication formulations.

We analyzed the distribution of concept and patterns of use separately for each domain. For conditions, measurements, procedures and observations we used the codes as they appeared in the data sources; for drugs we aggregated the codes to ingredient level using RxNorm hierarchy to account for different international drug formulations.

Data source granularity

For the purpose of quantifying granularity, we focused on Condition domain, which is the domain with the most comprehensive crosswalks between the source vocabularies (such as Read or ICD-9 codes) and standard SNOMED-CT ontology and is present in all of the data sets regardless of the provenance and country of origin [164].

To assess data source granularity, we introduced the term ‘granularity score’, which refers to the overall level of granularity of conditions in a data source and can be used as a relative metric to compare different data instances. We calculated the granularity score for each data source using three approaches described below. In each approach, we calculated the minimal number of steps (‘Is a’ relationships) within the SNOMED-CT hierarchy needed to get from a concept A to a concept B. These steps or levels of separation were used as a proxy for granularity, assuming that concepts within one level of separation have similar semantic distance to a parent (ancestor) term.

Three approaches (Figure 4, Step 1) differed in the ancestor from which to measure the granularity score. The reasoning behind using anchors (ancestor codes) was to have a consistent metrics for concepts and, therefore, for different concepts to be comparable.

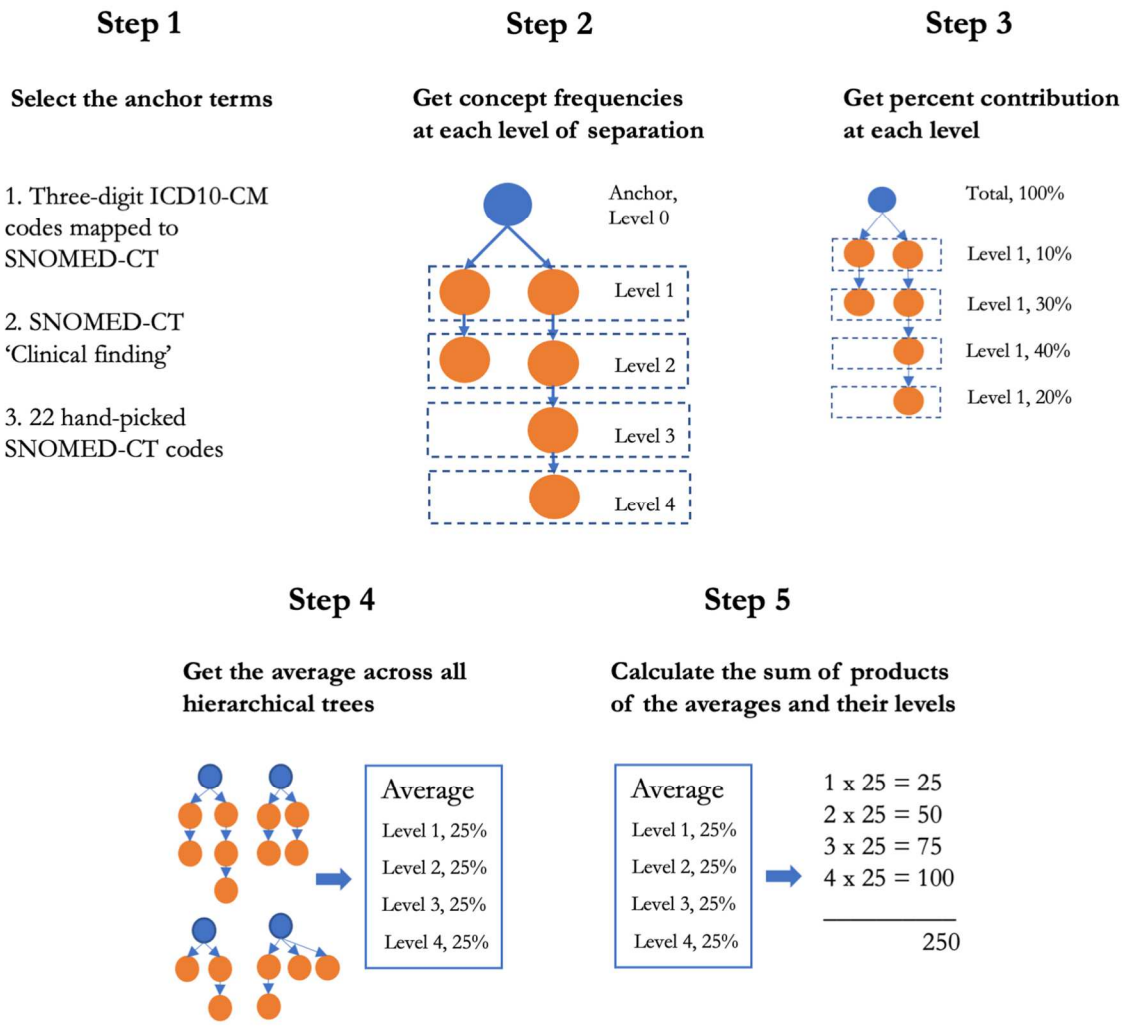


Figure 4. Overall study design for estimating condition-based data source.

We tested the following anchor concepts:

Approach 1. SNOMED-CT concepts mapped from three-character ICD10CM codes, excluding chapters 18-21 (signs and symptoms, injuries, external causes of morbidity and factors influencing health status). The purpose of this approach was to adjust for the fact that SNOMED-CT may have

different levels of granularity in different parts of the hierarchy, whereas the ICD10CM three-character codes may be less variable.

Approach 2. Broadest SNOMED-CT code 'Clinical Finding'. This assumes that in different parts of the hierarchy, the same degree of detail is encoded at about the same level down from 'Clinical Finding' for different diseases.

Approach 3. A set of 22 hand-selected SNOMED-CT codes that represent groups of conditions central to medicine (Table 2). In this way, we could manually ensure that the concepts were at a similar level of granularity.

Table 2. Twenty-two hand-selected SNOMED-CT terms that represent groups of conditions central to medicine, used as ancestor terms for calculating data source granularity scores.

SNOMED code	SNOMED category name	ICD10-CM group	ICD10-CM group name
55342001	Neoplastic disease	C00-D49	Neoplasms
362971004	Disorder of lymphatic system	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
111590001	Disorder of lymphoid system		
362970003	Disorder of hemostatic system		
299691001	Finding of blood, lymphatics and immune system		
362969004	Disorder of endocrine system		
74732009	Mental disorder	E00-E89	Endocrine, nutritional and metabolic diseases

118940003	Disorder of nervous system	F01-F99	Mental, Behavioral and Neurodevelopmental disorders
128127008	Visual system disorder	G00-G99	Diseases of the nervous system
362966006	Disorder of auditory system	H00-H59	Diseases of the eye and adnexa
271983002	Disorder of cardiac pacemaker system	H60-H95	Diseases of the ear and mastoid process
49601007	Disorder of cardiovascular system	I00-I99	Diseases of the circulatory system
50043002	Disorder of respiratory system		
53619000	Disorder of digestive system	J00-J99	Diseases of the respiratory system
80659006	Disorder of skin and/or subcutaneous tissue	K00-K95	Diseases of the digestive system
928000	Disorder of musculoskeletal system	L00-L99	Diseases of the skin and subcutaneous tissue
42030000	Disorder of the genitourinary system	M00-M99	Diseases of the musculoskeletal system and connective tissue
362972006	Disorder of labor / delivery	N00-N99	Diseases of the genitourinary system
173300003	Disorder of pregnancy	O00-O9A	Pregnancy, childbirth and the puerperium
362973001	Disorder of puerperium		
414025005	Disorder of fetus or newborn		

66091009	Congenital disease	P00-P96	Certain conditions originating in the perinatal period

For each of these anchor concepts, we obtained the frequencies of all the descendant concepts in the SNOMED-CT hierarchy at each level (Figure 4, Step 2) and calculated the distribution of concepts across different levels (Figure 4, Step 3). We then calculated the average distribution across all anchors (Figure 4, Step 4) and multiplied it with the corresponding levels of separation to arrive at a weighted distribution. Finally, granularity score was defined as the sum of weighted distribution obtained at Step 4 (Figure 4, Step 5).

Full process is described as:

$$\sum_{L \in 0,\ldots,max\ levels} \left(\frac{\sum_{a \in A} \left(\frac{\sum_{l:\, C_{l\ in\ L\ levels\ from\ C_a}} C_l}{\sum_{l:\, C_{l\ is\ descendant\ of\ C_a}} C_l} \right)}{N_A} \times L \right) \quad (1)$$

where C is the frequency of a concept, level is the level of separation, A is the set of ancestors (anchor codes) and N_A – number of ancestors.

Vocabulary granularity

Separately, we examined the granularity of vocabularies used in participating data sources (rather than the data sources themselves) to distinguish influence of different source coding schemas driving the granularity score as opposed to coding practices. To achieve that, we calculated the weighted distribution of target SNOMED-CT concepts across different levels of separation, where the levels were computed from three separate anchor codes described above.

Granularity applied to the real-world phenotyping tasks

Finally, to illustrate how granularity can be used to analyze applicability of data sources for phenotyping of specific disorders, we examined the granularity of databases for chronic kidney disorder. The most common code set for chronic kidney disorder (all concept in groups ICD-9(CM) 585 or ICD-10(CM) N18 'Chronic kidney disease') [239] is mapped to descendants of SNOMED-CT 709044004 'Chronic kidney disease' in the OHDSI Standardized Vocabularies. We took this code with all descendants and calculated total frequency of concepts at each level of the hierarchical tree.

3.1.3 Results

Code heterogeneity and utilization

The final dataset contained 272 billion records from twenty-two data sources: 14 US and 8 international (non-US). The data originated mainly from administrative claims (n=8), hospital charge data (3) and electronic health records collected in large teaching hospitals (5) or primary and secondary practices (6). The size of the datasets varied greatly, with the average number of 2.4 billion records (interquartile range, IQR 384 million – 17.8 billion) and 70.7 thousand (IQR 15.4– 102.4 thousand) unique standard concepts per data source. Full protocol with the data source description, total number of condition records and unique condition concept codes can be found on GitHub [242].

We observed high variability of the concepts used across participating data sources with a significant number of concepts (20%) to be found in only one dataset out of 22 (Figure 5).

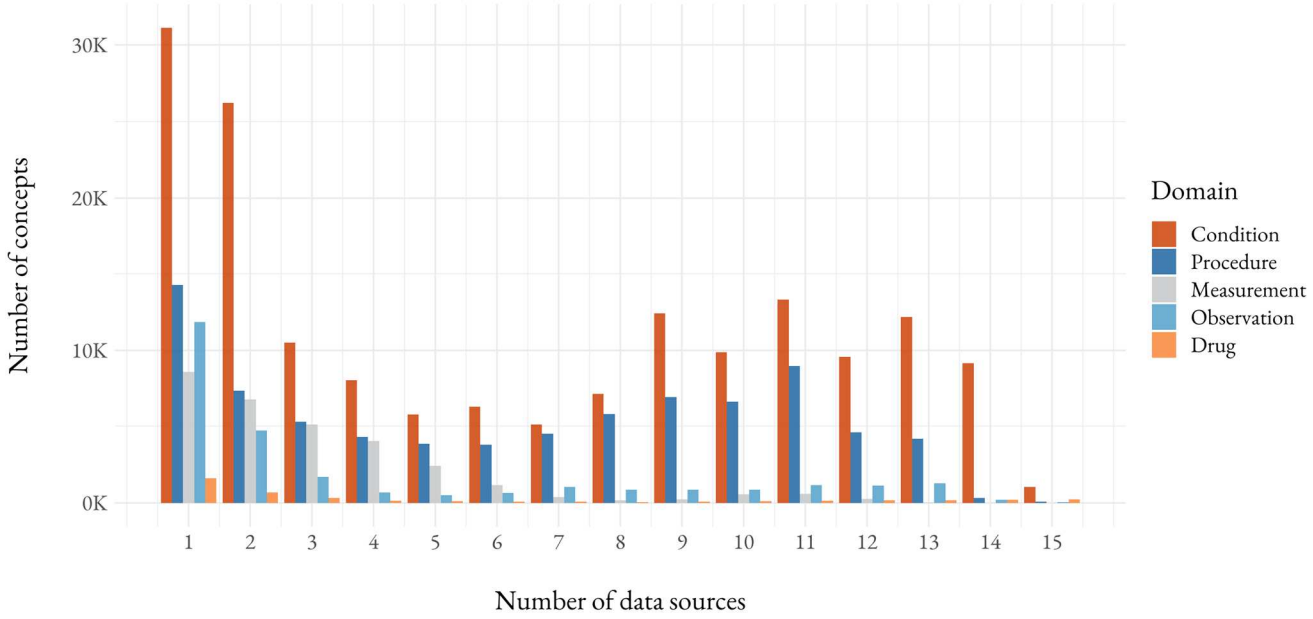


Figure 5. Number of concept codes per number of data sources they can be found in across the OHDSI network stratified by domain. Codes found in more than 15 data sources are omitted for visualization purposes.

Only 93 condition codes and none of the codes from other domains were found in all data sources. They were typically broad terms describing symptoms and states such as low back pain, ataxia and shock or common conditions such as anaphylaxis, angina pectoris or epilepsy.

Condition was the least heterogeneous domain with the highest rate of overlap across all domains, followed by Procedure, Measurement and Drug. Even then, 28.4% of measurement codes (such as lab tests and vitals) and 31.7% on drug codes on the ingredient level were unique to a data source.

While conditions and procedures are usually coded using a limited number of vocabularies, measurements and observations do not have established coding practices and are oftentimes coded as

either free text or using local terminologies, which may explain high heterogeneity across these domains. For example, myelin associated glycoprotein antibodies can be coded using 15 different codes depending on granularity of information available. Additionally, there are specific tests that are rarely performed, such as JAK2 gene exon 12 targeted mutation analysis in bone marrow.

Drugs only found in one dataset included herbal preparations, rarely used drugs such as ajmalicine, moxaverine or barbexaclone and specialized medications such as oxypertine.

Figure 6 shows that there was no apparent correlation between the frequency of the concepts in the datasets and the number of the databases they can be found in. For example, SNOMED-CT concept "Temperature" or LOINC measurement "Oxygen [Partial pressure] in Blood" can only be found in two datasets despite their high prevalence in those datasets. Nevertheless, most of the highly prevalent codes were found in six and more datasets, which shows that the datasets share widespread conditions, procedures and drugs.

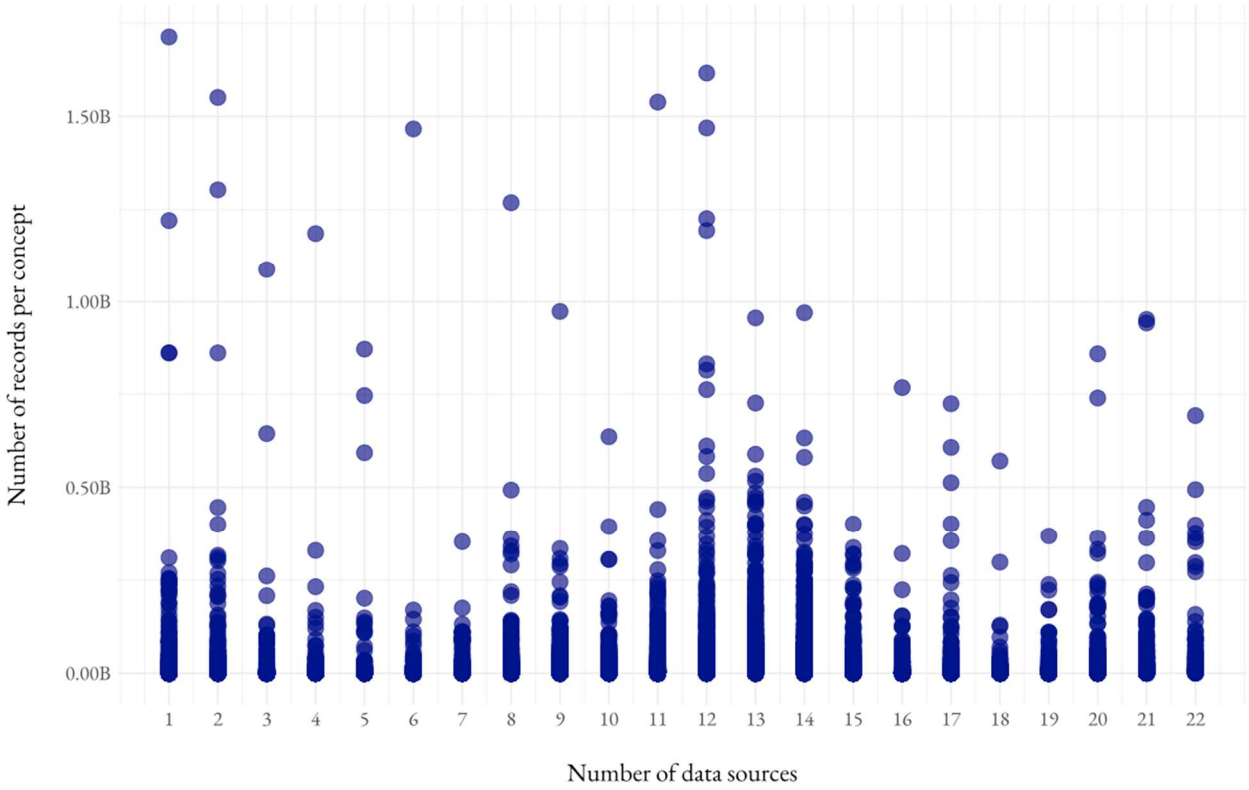


Figure 6. Frequency of the concept codes per number of data sources they can be found in across the OHDSI network. Each blue dot represents a concept code.

Data source granularity

We analyzed data source granularity using the three approaches and established 5 empirical granularity levels based on the distribution of granularities of the data sources: high, high/moderate, moderate, moderate/low and low. In most cases, all three approaches agreed (Table 3). For high/moderate and moderate/low data sources two approaches showed moderate granularity and one – high or low respectively.

Table 3. Granularity scores for selected data sources in the OHDSI network.

Database	Country	Type	Granularity score	

			by approach			Empirical level of granularity
			1	2	3	
AU-ePBRN	Australia	EHR	157	512	344	High granularity
Ajou University	Korea	EHR	117	516	347	High granularity
CUIMC	USA	EHR	114	519	355	High granularity
MDCR	USA	Claims	114	519	357	High granularity
NHIS/NSC Korea	Korea	Claims	111	510	336	Moderate/high granularity
StaRR	USA	EHR	113	509	345	Moderate/high granularity
HCUP	USA	Hospital charge	125	498	346	Moderate granularity
PanTher	USA	EHR	107	503	324	Moderate granularity
PREMIER	USA	Hospital charge	110	496	332	Moderate granularity
MDCD	USA	Claims	111	490	333	Moderate granularity
Hospital CDM	USA	Hospital charge	111	503	335	Moderate granularity
CCAE	USA	Claims	110	500	340	Moderate granularity
Open Claims	USA	Claims	110	505	342	Moderate granularity
Optum DOD	USA	Claims	110	506	342	Moderate granularity
Optum SES	USA	Claims	110	506	342	Moderate granularity
AmbEMR	USA	EHR	114	490	314	Moderate/low granularity
Tufts University	USA	EHR	118	477	331	Moderate/low granularity
DA France	France	EHR	100	490	304	Low granularity
DA Germany	Germany	EHR	100	472	309	Low granularity

JMDC	Japan	Claims	102	497	314	Low granularity
LPD Australia	Australia		112	475	311	Low granularity
MIMIC3	USA	EHR	178	474	343	Inconsistent granularity

Regardless of the approach, most of the data sources had moderate granularity (Figure 7). This group included mainly administrative claims (MDCD, CCAE, OpenClaims, OptumDOD, and OptumSES) and hospital charge data (Hospital, HCUP, and Premier) along with only one EHR source (PanTher).

High granularity data sources (AU-ePBRN, MDCR, CUIMC and Ajou University database) remained relatively granular regardless of the method used. StaRR and NHIS/NSC Korea appeared to be highly granular or moderately granular depending on the approach.

The low granularity group was the most homogeneous group, consisting of international data sources, which were primarily EHR-derived (LPD Australia, DA France and DA Germany), accompanied by one claims-derived source (JMDC). Another EHR source, AmbEMR, appeared as a low/moderate granularity data source.

Only one data source had noticeable inconsistency across approaches (MIMIC3) and was the smallest data source with only 749 condition codes.

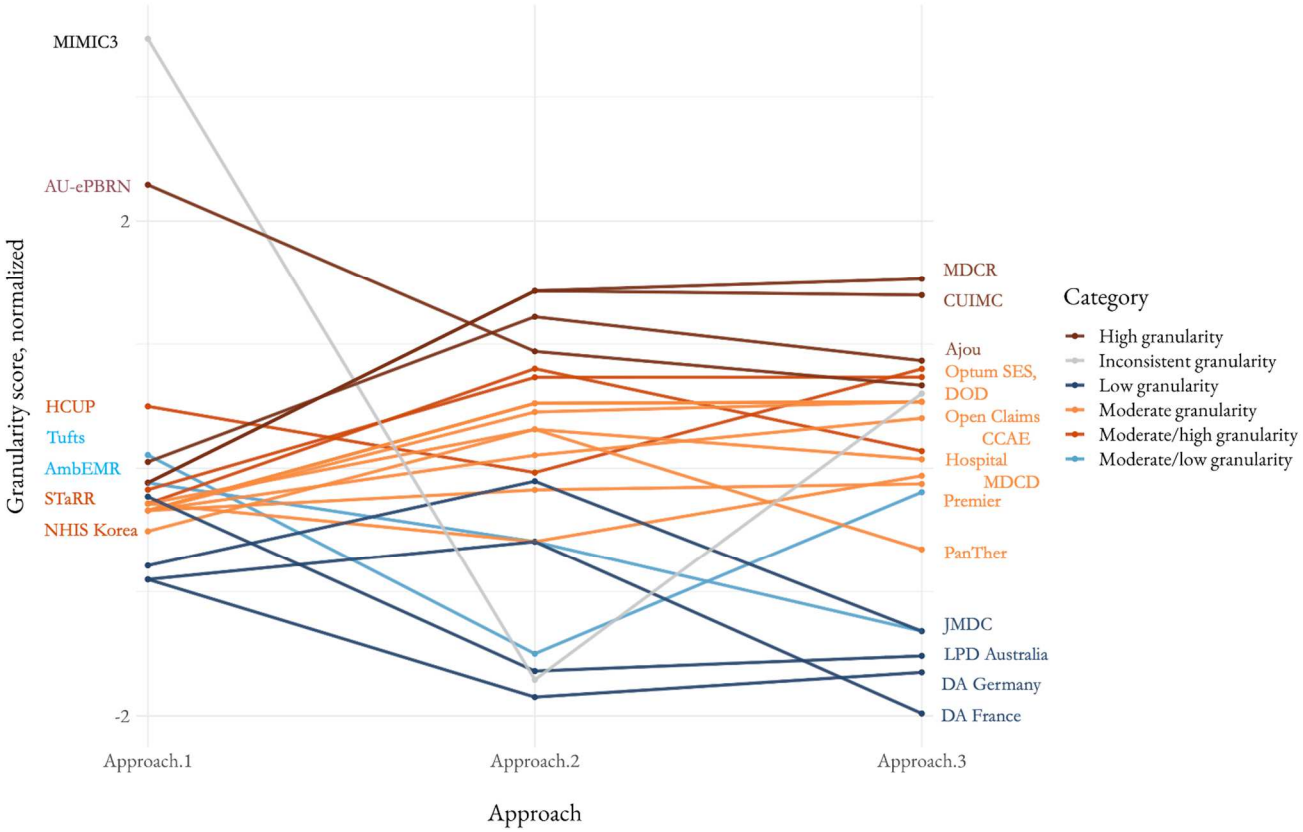


Figure 7. Normalized granularity score for the data sources across the OHDSI network stratified by granularity level.

We also found some patterns in data granularity related to data provenance. Overall, EHR data sources originating from primary and secondary care practices were less granular, while administrative claims data, hospital charge data and EHR data originating from large tertiary care hospitals were more granular. International data sources were on average less granular with only three out of eight non-US sources being moderately or highly granular (Figure 8).

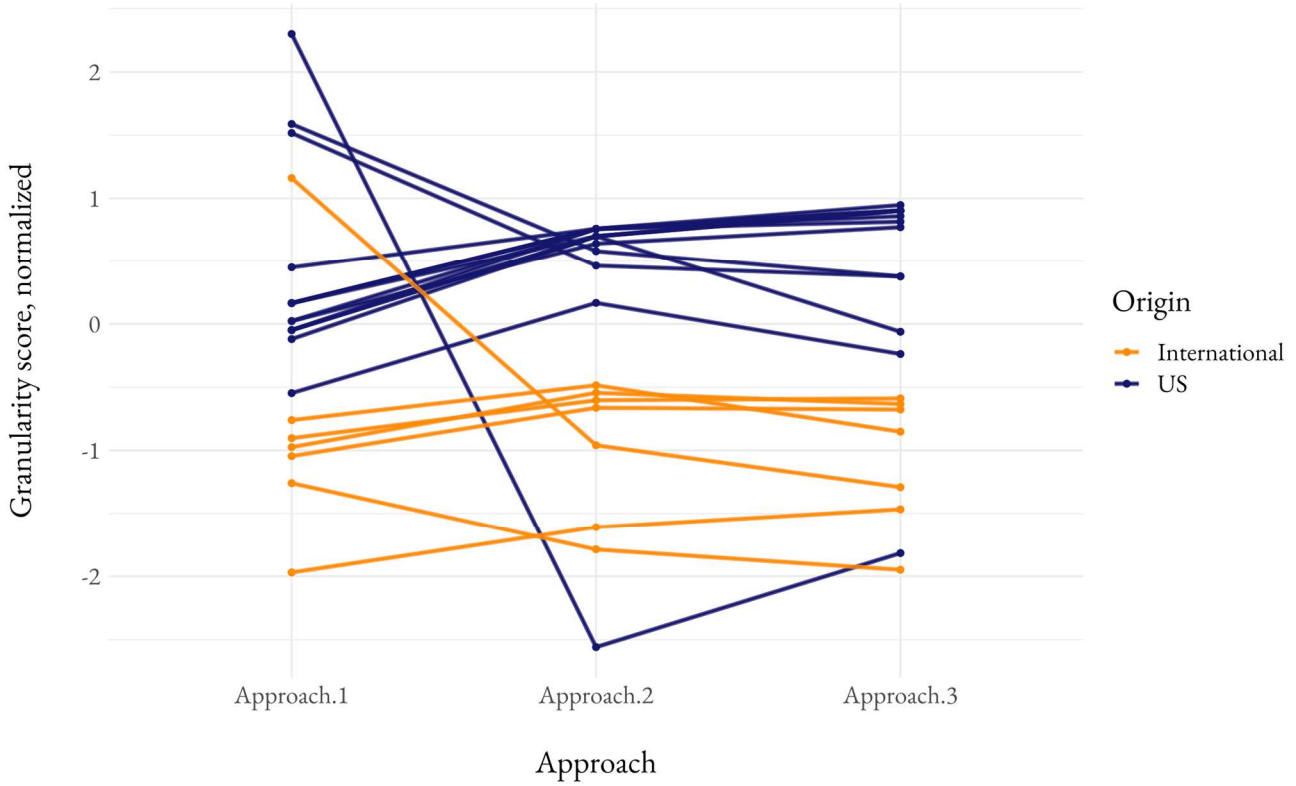


Figure 8. Normalized granularity score for the data sources across the OHDSI network stratified by the country of origin: US (blue) and international (orange).

Administrative claims data and hospital charge data showed similar patterns of granularity, but relative granularity within this group differed depending on the approach. MDCR had the highest granularity among other claims data, Optum DOD, Optum SES and OpenClaims had similarly moderate granularity, and MDCD with Premier had consistently low granularity in the group.

Granularity for specific disorders

Table 4 shows that on average 59% percent of records for chronic kidney disorder had relatively low granularity (condition with one attribute, such as 'Chronic kidney failure stage 3'). Some of sources comprised broader terms. For example, less precise concept 'Renal impairment' accounted for 23% of

all concepts related to chronic kidney disorder in LPD Australia. Given its prevalence, we may choose to examine the patients who have this code in their record to determine if they should be included in the cohort.

Table 4. Selected datasets for assessing granularity of chronic kidney disorder codes used across the OHDSI network.

Level of separation (example)	DA France	JMDC	LPD Australia	AmbEMR	CUIMC	MDCD	Average*
0 (Renal impairment			23.0%	4.0%	<0.1%		2.2%
1 (Chronic kidney disease)	94.5%	90.0%	32.0%	17.0%	25.4%	13.1%	24.7%
2 (Chronic kidney disease stage 3)	5.4%	8.9%	45.0%	68.8%	64.4%	67.4%	59.4%
3 (Chronic kidney disease stage 3 due to hypertension)	<0.1%	1%		10.0%	9.3%	18.5%	13.1%
4 (Malignant hypertensive chronic kidney disease stage 3)	<0.1%	0.1%		0.2%	1.0%	1.0%	0.5%
5 (Malignant hypertensive end stage renal disease on dialysis)				<0.1%			<0.1%

*Average frequency of concepts at a level across all data sources

3.1.4 Discussion

In this study, we explored code utilization and data source granularity across disparate data sources with different provenance of data, country of origin and various coding methods.

Variability and heterogeneity showed here as well as for the other distributed networks have to be accounted for in any observational study as failure to do so can significantly bias the results in any direction [105,107,149].

First, code sets developed on one data source are not likely to be generalizable to other data sources. Variability in code utilization remains even after data structure and content harmonization and reflects different coding practices, specifics of data capture, data cardinality and more. As we demonstrated, the data sources with similar provenance or country of origin may share some of the code utilization patterns but remain different enough to limit portability.

Second, as similar patients can be coded with various granularity in different data sources, it is important to be aware of the overall data source granularity to make informed decisions about phenotyping algorithms.

When using data sources with low granularity (as LPD Australia in this study), using less precise broad concepts is needed in order not to lose patients of interest. For example, when identifying patients with chronic kidney failure, researchers may opt for looking at broader concepts such as renal impairment. The latter accounts for nearly a quarter of all kidney disorder-related records in LPD Australia. Given that such a broad concept is not likely to initially be included in the phenotyping algorithm, it is important to recognize the fact that a large fraction of patients has this code.

We will now discuss three factors that contribute to data source granularity: (a) vocabulary, (b) population and (c) data capture process.

Vocabulary

SNOMED-CT is among the most comprehensive reference terminology available and is a mandatory standard vocabulary for conditions in the OHDSI network. SNOMED-CT supports polyhierarchy, where a concept may have multiple ancestors and inherits their meaning. Such polyhierarchies coexist in SNOMED-CT equally, so a single main hierarchical path cannot be identified. A concept can appear in multiple hierarchical trees at different levels, which obstructs assessing its complexity level when multiple anchoring codes are used. For example, 51292008 'Hepatorenal syndrome' appears in two hierarchical trees: 42030000 'Disorder of the genitourinary system' (5 levels of separation) and 53619000 'Disorder of digestive system' (2 levels of separation). While it poses challenges to establishing hierarchy-based granularity of an individual concept or an individual data source, such ambiguity is leveled out when aggregating across multiple codes and seeking relative comparison. Moreover, other ontologies such as ICD-10(CM) only support broad grouping of terms and not true polierarchies, which makes them unusable for estimating granularity of disparate data sources with different local ontologies.

We used different approaches that vary in anchor. Using top code 'Clinical Finding' as a single ancestor prevented duplication of codes across hierarchical trees and allowed us to capture all condition codes in the data sources. A disadvantage of such approach is participation of the concepts that carry insignificant clinical meaning. For example, ICD- 9(CM) concept 780.99 'Other general symptoms' frequently occurred in some of the data sources and, being mapped to SNOMED-CT 365860008 'General clinical state finding', conveyed little clinical meaning. Even if such a concept is present in a data source, it cannot be acted upon: it communicates too little clinical meaning to define any disorder of interest.

The ICD-10(CM) code-based approach was motivated by selecting patients in observational research, which is typically performed by selecting appropriate ICD-10(CM) codes to define disease or state. Such an approach can be inefficient when international data sources or data sources with unstructured data are involved. Indeed, source vocabularies in non-US data sources were less granular since ICD-10(CM), used in the US, is more granular than ICD-10 used internationally. If a feasibility part of network studies is performed on a highly granular data source, too specific concepts may be selected for phenotyping, which will lead to the patient loss.

On the other hand, this approach neglects concepts broader than the selected ICD-10(CM) counterparts, which can be of a particular interest in low granular data sources. The SNOMED-CT concept set approach (approach 3) overcomes this shortcoming by querying broad disorder groups.

In approach 2 and 3 duplication of concepts across different trees was offset by averaging those trees. Nevertheless, this approach will be sensitive to concept selection if concept space in the data source is limited. Although we tried to minimize this effect by excluding groups of disorders that have high overlap, duplicates can still be found and can potentially bias the granularity score for small data sources.

Population

Granularity can reflect the features of the population that had given rise to a data source. Unbalanced data sources with a focus on a specific population may be biased towards higher granularity for this population but remain otherwise non-granular. For example, 85% of MDCR patients are elderly, who tend to have more co-morbidities compared to young healthy patients [243]. Co-morbidities, in turn, are coded as granular complex concepts that reflect associations between disorders, e.g. 422166005 ‘Peripheral circulatory disorder associated with type 2 diabetes mellitus’ or 19034001

‘Hyperparathyroidism due to renal insufficiency’. Such high granularity is attributable rather to characteristics of the population (patients) than to characteristics of processes (data collection, coding or transformation). If a certain level of granularity belongs only to a specific portion of the data source, we need to disentangle this effect to be able to assess the baseline level of granularity. The latter will then reflect the granularity for the other groups of patients in a source, which can also be used for research.

We proposed to offset the influence of a particular patient group on data source granularity by stratifying concepts by disorder group (approach 1 and 3). In particular, it resulted in a reduced difference in the granularity of MDCD and MDCR, which was more extreme in approach 2.

Data capture process

The data can be generated to address different needs: electronic health records facilitate clinical records storage and retrieval, and administrative claims data are used in the reimbursement process. Clinical documents within electronic health records and administrative claims may capture similar patients differently. EHRs may tend to be less granular due to the nature of clinical workflow, while claims data can be more granular to maximize reimbursement.

It is supported by our observations that administrative claims data and hospital charge data were on average more granular than EHR data, especially if a data source originated from primary or specialty practices. Large hospitals’ EHR data appeared to be highly granular, which may suggest shared coding patterns.

We previously discussed granularity should be adjusted if a subset of patients influences granularity. Patient characteristics can also be viewed as a feature of data source granularity if the patient population is homogeneous. In this way, granularity has the potential to remain stable regardless of a selected fraction of patients.

Coding methods applied to unstructured data can also contribute to concept heterogeneity. Extracting data from clinical notes is a tedious and complicated process, which may decrease concept granularity as free text, especially if in large volume, may be converted to broad and imprecise structured data [244].

3.1.5 Limitations

We did not perform targeted SNOMED auditing to identify hierarchy inconsistencies, incomplete modelling or other issues described elsewhere [245,246]. As SNOMED is the most comprehensive and continuously growing reference terminology, we assumed that such issues will not be detrimental to assessing granularity or will influence all data sources equally.

In this study, we only analyzed conditions as a comprehensive hierarchy for procedures or measurements is lacking; including other domains in granularity score may be included in future work.

SNOMED-CT defines its concepts not only with hierarchical links, which we used in this study, but also with 'has-a' relationships, which can potentially be used to assess granularity. While attribute-based granularity inference is complicated by inconsistencies in assigning attributes and high volume of relationship types [245], future work may include comparing hierarchy-based approaches to attribute-based approaches.

3.2. Recommender system for comprehensive concept set creation[2]

In the previous section we observed that data sources are highly heterogeneous and have various levels of granularity, which makes code sets developed in one institution not likely to be generalizable to the other sites unless the researchers account for heterogeneity. The issue of creating comprehensive concept sets that contain all appropriate codes remains largely unsolved in the setting of moving learned phenotypes among disparate institutions without retraining.

In this section, we discuss how we can use the dataset we collected across the network in a recommender system for creating comprehensive concept sets. I use mixed-method approaches to pre-compute recommendations for all codes in the OHDSI Standardized Vocabularies, and then develop, deploy, and evaluate an open-source recommendation system. I find that creating cohorts with the system allows to identify substantially more patients while preserving positive predictive value and capture patients early in the course of the disease to reduce index event misspecification. I follow up by showing how it integrates into the OHDSI pipeline for phenotype development, evaluation and storage and demonstrate its utility as used in 11 major OHDSI network studies.

3.1.1 Background

As we discussed previously, concept selection, either expert-based or data-driven, plays an essential role in phenotype algorithm development and subsequent patient selection.

---

[2] This section is published in the AMIA Annual Proceedings 2022. The full citation for this publication is Ostropolets A, Ryan P, Hripcsak G. Phenotyping in distributed data networks: selecting the right codes for the right patients. *AMIA Annual Proceedings* (2022).

While inaccurate concept sets (incomplete or those that contain irrelevant concepts) concept sets may introduce bias and shift the study estimates in any direction, there is a lack of research on systematic approaches to building comprehensive yet accurate concept sets.

Common expert-based approaches for concept selection include deriving codes from published phenotypes or using local expert knowledge [166,238]. When we looked at a random subset of 40 papers published in top clinical journals [247], we observed that a substantial number of papers rely on the previously developed concept sets, even if the referenced papers did not specify performance metrics. We can hypothesize that (a) peer-reviewed publication is enough to establish trust in concept sets and (b) the process of concept selection is oftentimes time-consuming, ambiguous, and unclear. Indeed, it is rarely formalized, and the researchers often turn to clinicians to learn about meaningful concepts.

While useful from the clinical stand point, local expert knowledge may fail to produce a complete set of codes due to high variation in billing and coding practices among providers [107].

As the process of concept selection is rarely described, the reasoning behind code inclusion or exclusion is mostly non-reproducible and hard to assess [248]. This challenge is especially complex in networks, where the analysis is performed across the institutions in different countries, coding practices and ontologies. In that case, the high volume of concepts makes the process of searching for relevant concepts almost non-feasible. For example, ICD-10(CM) and ICD-9(CM) alone contain more than 113 thousand codes and the number grows rapidly with concepts from international ontologies.

Another common approach is to use machine learning approaches like multi-view banded spectral clustering [174], non-negative tensor-factorization [175], latent Dirichlet allocation [249], embeddings with knowledge graphs and EHR [177] to derive concepts from the structured data or supplement the latter with unstructured data [178,179]. While it potentially eliminates human

subjectivity, learned knowledge is still limited to the institution the model was developed in and therefore may not be applicable to other institutions [109].

The process of code selection in networks with multiple institutions is not standardized, which leads to variable and non-reproducible approaches within individual institutions and potentially biases the results of the studies. Moreover, it greatly delays evidence generation. Current recommendations suggest extensive data exploration, inspection of the documents on coding practices and discussions with data owners to create accurate concept sets [250]. We propose that the data set of concept use across the network can be used to guide concept selection to achieve (a) systematic and (b) fast and feasible approach to code selection even when the researcher does not have the access to all participating data sources.

3.1.2 Methods

This section describes three parts of the study: (1) methods to identify similar concepts, (2) a recommender system for building comprehensive concept sets and (3) system evaluation.

Methods to identify relevant concepts

We operated with the data set described in the previous section, which consisted of eleven million unique concepts appearing in at least one source within the network, with 272 billion records summarized.

For each concept we calculated an aggregated estimate of the frequency of its use across all the data sources as well as an aggregated estimate of the frequency of use of all its descendants. The latter was derived using the OHDSI Standardized Vocabularies and represented the aggregated frequency of all child concepts in a corresponding hierarchy (RxNorm, SNOMED, LOINC etc). For the concepts that

do not participate in the OHDSI Standardized Vocabularies hierarchy (such as ICD-10(CM), Read or NDC), the estimate of descendants' frequency was equal to the aggregated frequency of the code itself.

Aggregated frequencies were used to pre-compute a set of recommended terms for each standard code in the OHDSI Standardized Vocabularies. We applied mixed-methods techniques to derive recommendations (Figure 9).

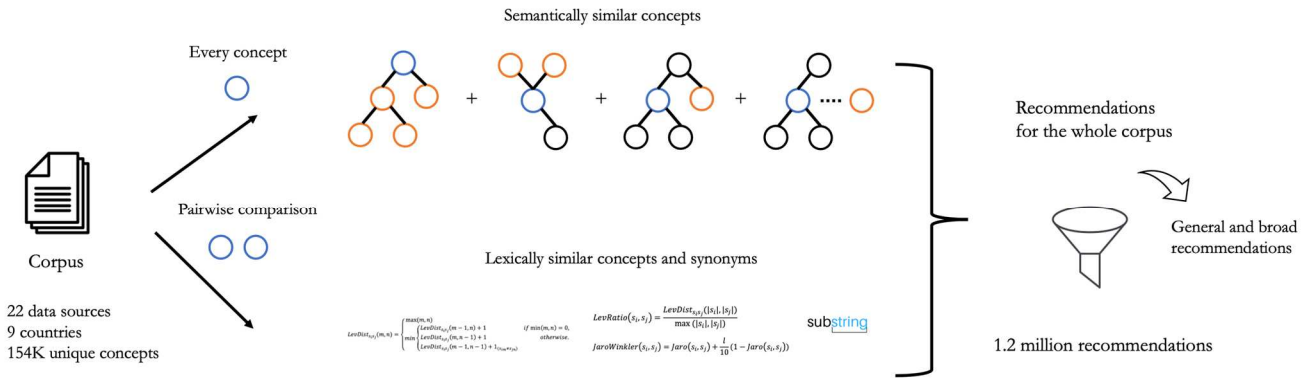


Figure 9. Overview of the methods used to derive recommendations for building comprehensive concept sets.

First, we found semantically similar concepts. They were added by selecting the most proximal concepts within a hierarchy and through the crosswalks between the adjacent ontologies such as HCPCS, CPT-4, SNOMED and ICD10PCS. These included both the concepts proximal and distal to the common ancestor as well as those belonging to a different hierarchy sub-tree (no common ancestor).

Then, we found lexically similar concepts. We leveraged both concept names and concept synonyms obtained from terminologies and ontologies (such as SNOMED-CT or CPT-4).

For each concept, we three lexical similarity metrics (Levenshtein distance, Levenshtein ration and Jaro-Winkler), which, along with substring matching were used to select a set of recommendation candidates in the corpus [251].

1. Levenshtein distance, which is a recursive definition for the absolute Levenshtein distance between two strings:

$$LevDist_{s_i s_j}(m,n) = \begin{cases} \max(m,n) \\ min \begin{cases} LevDist_{s_i s_j}(m-1,n)+1 & if\ \min(m,n)=0, \\ LevDist_{s_i s_j}(m,n-1)+1 & otherwise. \\ LevDist_{s_i s_j}(m-1,n-1)+1_{(s_{im} \neq s_{jn})} \end{cases} \end{cases}$$

where m and n index character positions in two compared strings.

2. Levenshtein ratio, which is normalized Levenshtein distance using the maximum length of each string:

$$LevRatio(s_i, s_j) = \frac{LevDist_{s_i s_j}(|s_i|,|s_j|)}{\max(|s_i|,|s_j|)}$$

3. Jaro-Winkler, which is a modification of Jaro metric giving higher weight to strings that match beginning at a set prefix proportional length:

$$JaroWinkler(s_i, s_j) = Jaro(s_i, s_j) + \frac{l}{10}(1 - Jaro(s_i, s_j))$$

where

$$Jaro(s_i, s_j) = \begin{cases} 0 & if\ m = 0, \\ \frac{1}{3}\left(\frac{m}{|s_i|} + \frac{m}{|s_j|} + \frac{m-t}{m}\right) & otherwise. \end{cases}$$

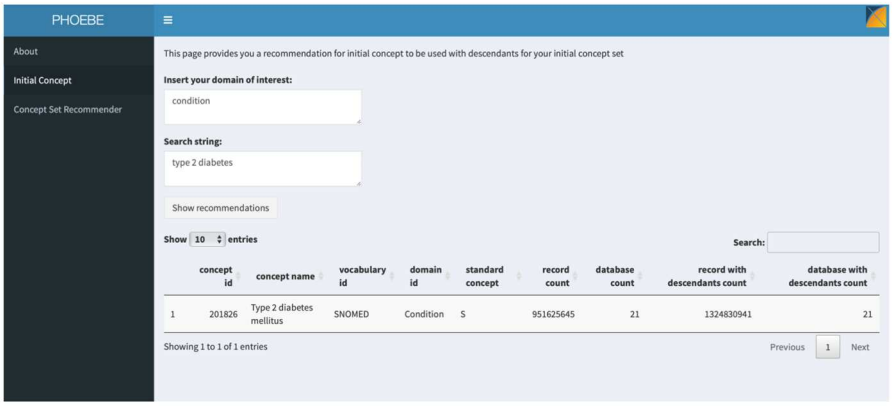
The concepts were subsequently filtered to those that provide an adequate data source capture. That was achieved by eliminating the concepts from those ontologies that cover less than half of the data

sources (either directly or through mapping from the source ontologies) in a given domain. We also removed the concepts not used in any data source. Finally, we removed the broad terms that are not likely to contribute to a phenotype definition (such as ‘Disorder’ or ‘Family History of Disease’).

The final set has all code-similar code pairs identified across all techniques.

Recommender system

These pairs were used by our recommender system – Phenotype Observed Entity Baseline Endorsements (PHOEBE). We created an R Shiny-based application available publicly (https://data.ohdsi.org/PHOEBE), which has two parts: initial concept recommender and concept set recommender (Figure 10).



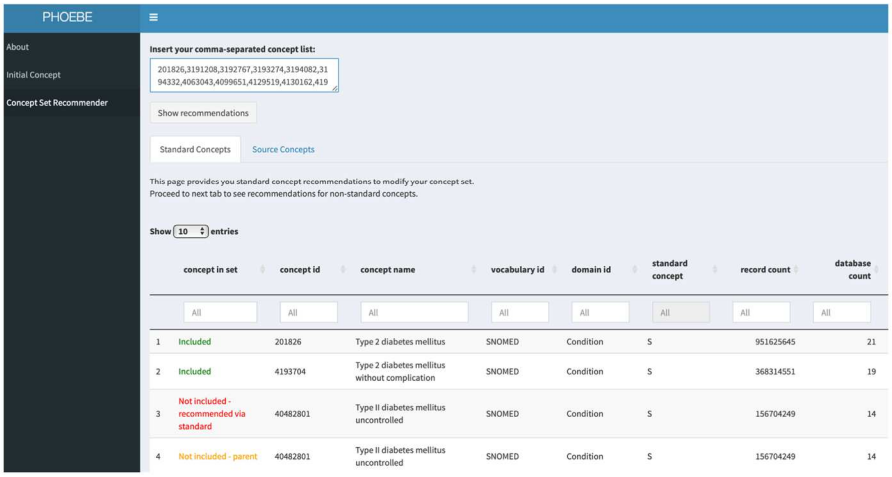


Figure 10. PHOEBE R Shiny application: initial concept and concept set recommender.

In the initial concept tab, a user inputs a string that specifies her clinical idea (such as diabetes mellitus type 2) and PHOEBE outputs the best match for a code that represents this clinical idea prioritized based on the number of data sources covered and an aggregated frequency of both the code itself and its descendants. The user can then use this code with all its descendants as a starting place for developing the concept set.

Concept set recommender takes a string of codes (obtained in the previous step) and outputs a set of codes divided into the following groups:

- included codes

- not included codes (ancestor and descendant codes of the codes included in the concept set)

- recommended codes (recommended through matching in source vocabularies or standard vocabularies).

The output is prioritized based on the aggregated frequency to focus decision-making on the concepts that provide the largest gain in record count.

Evaluation

For validation, we used an electronic health record data source (CUIMC EHR) and three administrative claims data sources (CCMR, MDCD, MDCR) translated to OMOP CDM (Table 1). We selected diabetes mellitus type II, diabetes mellitus type I, heart failure and attention deficit hyperactivity disorder (ADHD) – common conditions that were extensively studied in the observational literature and for which drug treatment exists. For each of the conditions, PHOEBE was used to create concept sets following the steps outlined above.

As a reference, we selected two eMERGE network phenotypes representing the same clinical ideas [151]. eMERGE is a national network for high-throughput genetic research that developed and deployed numerous electronic phenotype algorithms. Phenotypes have been created by highly qualified multidisciplinary teams and often been taking up to 6–8 months to develop and validate. For each eMERGE phenotype, we extracted ICD-9(CM) concept sets used in the original implementation and translated them to SNOMED-CT concepts [6]. We then created patient cohorts by selecting patients with at least one occurrence of a diagnosis code from corresponding concept sets and with at least 365 days of prior observation to ensure data coverage. For each cohort, we followed patients to look for a specific treatment, which included any occurrence of insulin products for type I diabetes mellitus; oral antidiabetic drugs (metformin, sulfonylureas, thiazolidinediones, dipeptidyl peptidase IV inhibitors and glucagon-like peptide-1 agonists) for type II diabetes mellitus; beta blockers, angiotensin-converting enzyme inhibitors and diuretics for heart failure and viloxazine, atomoxetine, amphetamine methylphenidate and guanfacine for ADHD [252–255].

We computed the positive predictive value (PPV) of each phenotype, which was defined as a proportion of people with a diagnostic code who also had subsequent treatment with corresponding drugs. Additionally, for patients with subsequent treatment identified by both PHOEBE and eMERGE concept set-based algorithms, condition index dates (the date a disease was first observed in a patient) were extracted and compared.

3.1.3 Results

Recommendations

Overall, we generated more than 1.2 million recommendations with on average 3 recommendations per a code (interquartile range [IQR] 1-12). Most of the recommendations were generated for conditions, drugs (different formulations) and procedures.

Most of the recommendations were within the same domain. For example, for diabetes mellitus type PHOEBE recommended SNOMED-CT codes that are relevant to diabetes but are in different places in SNOMED-CT hierarchy: “Hyperosmolar coma due to type 1 diabetes mellitus”, “Type 1 diabetes mellitus uncontrolled “, “Peripheral circulatory disorder due to type 1 diabetes mellitus”, “Disorder of nervous system due to type 1 diabetes mellitus” and other complications. This allows to capture not only all ICD-10(CM) and ICD-9(CM) codes mapped to SNOMED-CT but also Read codes and other international terminologies.

For diabetes mellitus type II, PHOEBE also recommended complications of diabetes along with broad terms such as “Diabetic – poor control”. While the latter concept is broad, it accounts for 833,654 records in the OHDSI network. Given higher prevalence of type II diabetes compared to type I and high utilization of this code, researchers may want to use this code in their phenotype with additional restrictions like older age, absence of codes for type I diabetes or others.

Cross-domain recommendations were also generated. For example, codes relevant to abscess of appendix include various procedures performed on abscess such as LOINC “Guidance for drainage of abscess and placement of drainage catheter of Appendix”, ICD-9Proc “Drainage of appendiceal abscess” or CPT-4 “Incision and drainage of appendiceal abscess, open”.

Phenotyping pipeline

The main utility of PHOEBE lies within recommending lexically and ontologically similar concepts and allowing to inspect code sets based on real-world code utilization in the network.

As a result, over the past two years, we participated in the numerous clinical studies where we built more than 170 phenotypes for conditions like diabetes type 2, acute kidney failure, chronic obstructive pulmonary disease, rheumatoid arthritis, chronic heart failure, venous thrombosis, epilepsy, Guillain-Barre syndrome and many more [17,19,34].

Since 2020, PHOEBE has been used in eleven major network studies conducted in the OHDSI network, both led by us and by other researchers. They include clinical studies such as characterizing patients with COVID-19 [18,22,26,27,29,31,35,256,257] or studies of adverse events [23,33] and methodological studies such as investigating the sensitivity of background rates (further described in Section 3.1) or validation of prediction models [28,258]. PHOEBE is continuously used as a part of phenotyping pipeline for developing, evaluating and storing phenotypes (Figure 11) in the new studies like Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) for type 2 diabetes mellitus [21]. The studies were published in high-impact journals like BMJ, Nature Communications, Lancet Rheumatology, Journal of Asthma, Drug Safety, Rheumatology, BMJ Open and others.

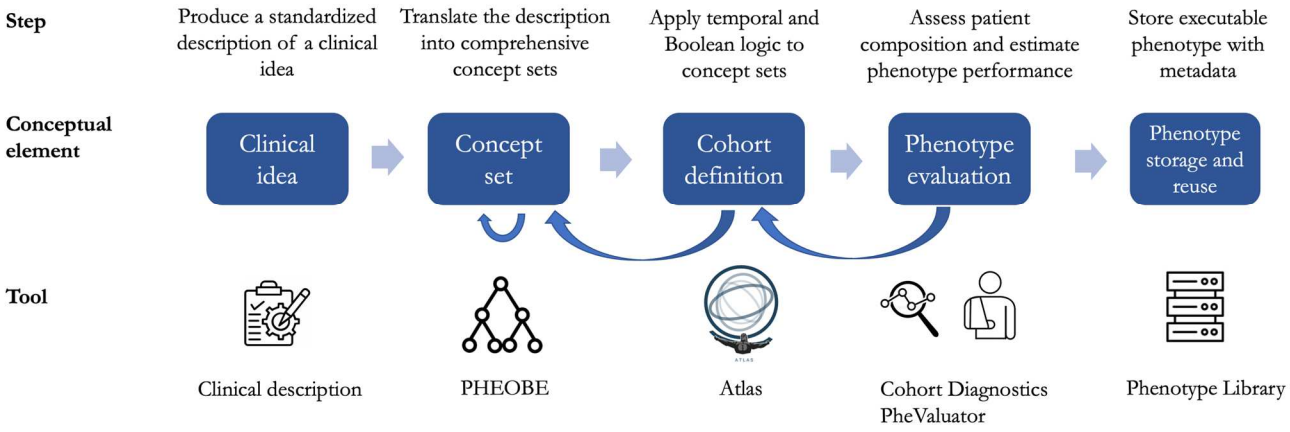


Figure 11. OHDSI pipeline for developing, evaluating, and storing phenotypes, including the OHDSI informatics tools (PHOEBE, ATLAS, Cohort Diagnostics and Phenotype Library).

Within this pipeline, PHOEBE is used to (a) consistently select a starting code for a code set that represent a clinical idea and (b) iterate over until a comprehensive code set is generated. After the logic is applied to code sets to create a cohort definition (typically in the OHDSI tool Atlas), the latter can be executed against a data source to create a cohort of patients. The aggregated characteristics of the cohort are them inspected (Cohort Diagnostics) to assess their plausibility, which can lead to further iterations over cohort definitions or code sets to achieve the patient composition that matches a hypothetical cohort of patients of interest.

Evaluation

While the main benefit of PHOEBE is providing the data to support researchers' decision making, we also formally evaluated its impact on patient selection.

When looking at the cohorts created using the concept sets provided by PHOEBE and used in the eMERGE phenotypes, we found that on average PHOEBE identified more patients while preserving similar positive predictive value (Table 5). We observed high heterogeneity in the magnitude of patient gain among data sources and conditions. For example, for type I diabetes, the algorithm created with PHOEBE identified approximately the same set of patients as the algorithm that used well-curated concept set from eMERGE. Notably, the cohort of type II diabetes patients created using PHOEBE had more than 5 times more patients and had a higher positive predictive value.

Table 5. Comparison of eMERGE and PHOEBE concept set-based algorithms' performance. PPV – positive predictive value.

		eMERGE cohort			**PHOEBE cohort**			**Cohort overlap**
		Patients with subsequent treatment	Total	PPV	Patients with subsequent treatment	Total	PPV	
Diabetes mellitus type I	CUIMC	7,599	25,701	0.34	7,659	25,884	0.34	25,635
	MDCD	25,954	191,710	0.13	26,240	194,100	0.13	189,810
	MDCR	19,274	185,874	0.10	18,684	181,712	0.11	179,236
Diabetes mellitus type II	CUIMC	42,200	211,904	0.20	44,828	218,127	0.21	211,903
	MDCD	342,975	1,807,688	0.19	383,435	1,849,177	0.21	1,807,019
	MDCR	105,449	386,265	0.27	804,415	2,172,925	0.37	386,237
ADHD	CUIMC	5,922	29,890	0.20	7,002	35,291	0.20	29,890

	MDCD	722,485	1,471,559	0.49	809,093	1,657,347	0.49	1,471,542
	MDCR	7,158	17,065	0.42	7,875	18,535	0.43	17,065
Heart Failure	CUIMC	71,281	162,297	0.44	74,121	168,111	0.44	162,281
	MDCD	332,929	870,080	0.38	342,112	893,807	0.38	869,500
	MDCR	892,689	1,166,980	0.76	919,077	1,204,488	0.76	1,165,900

In general, patient gain was less notable in CUIMC, which can be explained by the fact that eMERGE phenotypes were partially developed on CUIMC data. Even when there was no significant difference in PPV, PHOEBE's algorithm identified more patients with subsequent treatment.

We obtained similar performance upon repeating the procedure for other randomly selected versions of these phenotypes found in the literature.

Aside from evaluating the ability of PHOEBE to identify patients of interest, we also examined differences in the index dates (date of first observation of a disease in a data source) for the patients identified by both PHOEBE and eMERGE concept set-based algorithms. We observed that PHOEBE can identify patients earlier on in the course of the disease (Table 6).

Table 6. Comparison of condition onset date (index date) in patients with different index dates identified by both eMERGE and PHOEBE concept set-based algorithms.

	MDCD		**MCDR**		**CUIMC**	
	Patients, n	Difference in days, median (IQR)	Patients, n	Difference in days, median (IQR)	Patients, n	Difference in days, median (IQR)
Diabetes mellitus type I	2,243	371 (139-795)	2,807	407 (176 – 894)	206	1,001 (206-1,915)
Diabetes mellitus type II	253,728	77 (22-226)	957	122 (13-417)	11,536	189 (28-873)
ADHD	165,590	201 (59-525)	203	175 (61-318)	1,801	146 (35-532)
Heart Failure	16,941	49 (3-298)	25,294	107 (6-472)	2,242	78 (6-724)

For example, patients with different disease onset dates, on average, presented with diabetes mellitus type I more than two years earlier when using the PHOEBE algorithm in CUIMC and more than one year earlier in MDCD and MDCR. This can be explained by multiple patients having unspecified diabetes mellitus codes prior to being treated by endocrinologists. This pattern was consistently observed in all conditions and data sources, which may reflect that the use of broad non-specific codes by clinicians is common in early stages of disease. This finding suggests that the use of concept sets consisting of only narrow specific codes may introduce index even misspecification as such codes may capture patients not at the first observation of the disease but later on in the course.

3.1.4 Discussion

Selecting relevant codes for a phenotype is not a trivial task, especially when conducting a study on multiple data sources. Inclusion or exclusion of codes can introduce a significant bias in the study and influence study estimates in both directions.

One of the most common approaches is borrowing concept sets from the previously published studies. Yet, as widely acknowledged, the performance on one data source does not guarantee similar performance on another. For single-center studies, the researchers can get away with leveraging local data or expert knowledge to produce concept sets. The former approach was also advertised for organizing data models that do not involve semantic standardization, such as Sentinel or PCORNet [162,166]. While expert knowledge is critical in both phenotype development and evaluation, clinicians are oftentimes not familiar with the data or code use patterns [248]. While they may tend to extrapolate their practices to the other clinicians, real-world patterns of treatment are highly heterogeneous and variable [259].

Moreover, for published observational research the choices made for concept sets are rarely described. The readers can assess neither how the concept sets were arrived at, nor the implications of the choices made. We have no reason to believe that different institutions or individuals have the same approach to concept set selection, which means that a concept set, phenotype and, in turn, study results can be different when another researcher performs a selection for a given study. Using purely data-driven approaches with sufficient performance eliminates variability induced by an individual but provides little help when a study needs to be executed in a network. Moreover, the current large-scale

initiatives that have a potential to impact decision-making are limited to specific domains or conditions and take years to develop and implement locally [148,260].

An intermediate between expert-based and data-driven approaches is a data-augmented expert-based approach, which uses the data to guide decision-making. There are several studies on describing how the data was used to learn the patterns for code use and guide clinicians and informaticians in concept set selection [167,169]. While providing important insights, they are limited to a single center or condition of interest and do not provide general guidance on how to create a concept set. Moreover, they do not describe the process for international networks that use disjoint ontologies to code their data.

In this work, we showed how a recommender system can leverage the code usage patterns to provide clinicians and informaticians with the data to ground their code selection decisions. While the final choice of including or excluding a code in the concept set is left to the researchers' discretion, having the data allows them to inspect included and excluded codes prioritized based on their frequency of use. Such an approach, as shown by us here, can reduce measurement error and index event misspecification and can be used by the researchers outside of the OHDSI collaborative.

Having the data from the network also improves speed and scalability. The current recommendations for code selection emphasize the need for extensive review of the local coding practices and heavy engagement of the data owners [57], which substantially lengthens the initial steps of observational studies. PHOEBE partially alleviates this problem by providing the code utilization across different institutions and allows to partially standardize and formalize the process. Overall, based on our experience, the process of code selection becomes more efficient as all the contributors can assess code lists, recommendations and approximate the influence of each code on patient selection instead of browsing thousands of codes in the ontologies.

At the same time, a high demand for efficient code selection and phenotyping, which we observed in the network, highlights a need for further automation of the process of code selection. Potentially, patient characteristics of the resulting cohort can inform code selection.

3.2.5 Limitations

One of the limitations of this work is using only the US data sources for evaluation. As PHOEBE leverages data from a large network, it may show better performance on non-US data sources, which is yet to be shown. We used specific treatment as a proxy for patients being a true positive case. While imperfect, it allows comparing algorithm performance in absence of reliable methods to validate phenotypes in administrative claims data.

3.3. Impact of phenotype logic design choices on patient characteristics

As we observed in the previous sections, concept selection influences on patient selection. Once concept sets are defined, the next step that can introduce bias in the study is defining Boolean or temporal logic that is applied to the concept set to create an executable algorithm (cohort definition) that, when executed against a data source, translates into a cohort of study subjects. Design choices such as the order of inclusion and exclusion criteria and the specifics of how the criteria are constructed may influence patient selection, study estimates and validity.

This section focuses on quantifying the variability in translation of one conceptual definition into an executable algorithm and examining its impact on patient characteristics.

We conduct an experiment with 45 OHDSI collaborators, which was set up as a standardized implementation of the clinical description from a recent high-impact observational study. We execute all implementations and compare them to the master algorithm created together with the original investigator and observe high variation (up to 10-fold difference) in cohort size as well as in patient baseline characteristics.

3.1.1 Background

While it may seem obvious that variations in algorithmic implementations of a clinical idea influence patient selection and, potentially, study estimates, there is limited research on the magnitude of this effect. As previously acknowledged, phenotype development is a highly iterative process that, among other things, involves modifying inclusion and exclusion criteria to capture patients of interest and, if performance is evaluated, optimize sensitivity and specificity [151].

Yet, approaches for modifying and testing different combinations of criteria in a systematic and reproducible fashion do not exist. Most of the existing clinical literature does not report the details of the

process and whether more than one combination of inclusion and exclusion criteria was considered. Some studies reverse-engineer the rules based on diagnosis and treatment trajectories of already established cases [261]. Other create variations of phenotypes that are intended to be sensitive or specific and characterize the patients who are identified by either or both phenotypes [169].

Some (not all) studies report implementing and evaluating more than one version of clinical idea to create the best performing phenotype [19,108,262]. As we demonstrated in our brief review of observational studies published in top clinical journals, a large number of clinical studies re-use the phenotypes from previous literature or chose one phenotype without explicitly specifying why such logic and design choices were made [247]. In this way, variability coming from the deviations in implementation of clinical idea into an executable algorithm remains unmeasured and underappreciated.

Moreover, it is unclear if changes in implementation would lead to different point or interval estimates and, as the result, would modify the conclusions of the study.

In this study, we aimed to (a) assess how the study design described in a high-quality observational research study could be interpreted by multiple teams of independent researchers and (b) quantify the impact of the variability of operational logic choices made by those teams on patient characteristics.

3.1.2 Methods

To focus on variability stemming from different logic implementation we chose a paper that thoroughly describes a clinical idea and set up an experiment to implement the latter as an executable algorithm. We selected the article by Albogami et al. based on its robust methods, clinical relevance, and availability of the data source used, as well as the completeness of the study design description in the main body of the text and supplemental materials [263]. The study was published in 2021 and

investigated an association of glucagon-like peptide 1 receptor agonists (GLP-1RA) and chronic lower respiratory disease (CLRD) exacerbation in a population with type 2 diabetes mellitus (T2D) and CLRD.

Conceptual definition

The manuscripts' cohort definition was new adult users of GLP1-RA with diabetes mellitus and chronic obstructive lung disorder. This seemingly simple clinical idea had, in fact, multiple sub-criteria. Patients have to be at least 17 years old to be adults; GLP1-RA by default is an add-on therapy, so a patient has to be on a first-line antidiabetic drug as well; to ensure that a patient has type 2 diabetes they are required to have no prior insulin exposure or prior type 1 diabetes mellitus (to exclude type 1), not being pregnant (gestational diabetes) and not having conditions requiring corticosteroid treatment (secondary diabetes); to ensure that a patient has COPD they should not have cystic fibrosis, lung cancer, pulmonary embolism or pulmonary hypertension.

Master implementation

Based on the conceptual definition, together with the original author, we constructed a cohort definition using the OHDSI tool ATLAS (Figure 12). ATLAS is a web-based application that allows defining phenotypes, constructing and executing cohorts against local data source(s), characterizing subjects in a cohort and designing and implementing various observational studies [264]. The definition specified the entry event upon which a patient enters the cohort (first GLP1-RA exposure in 2007-2017), ten inclusion and exclusion criteria and the exit event upon which the patient leaves the cohort (is right-censored). Each inclusion and exclusion criterion comprised a start and end date, a duration (for drug exposures), one or multiple associated concept sets, a set of Boolean or temporal logic applied to the

concept set(s) and an order in which the criteria were applied. The master implementation used a list of pre-defined concept sets created in collaboration with the original author [265].

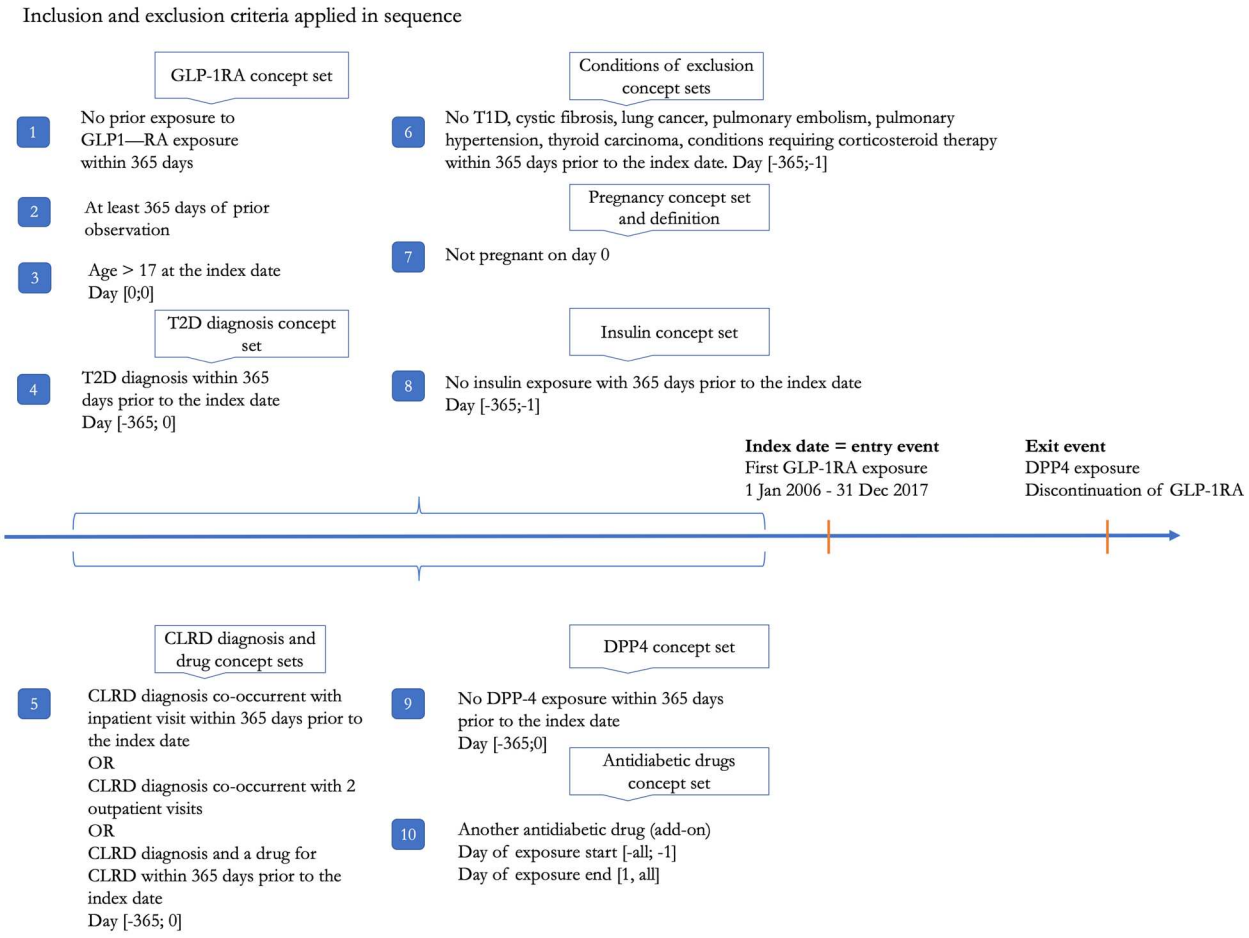


Figure 12. Master new GLP1-RA user cohort implementation: entry and exit event and 10 inclusion and exclusion criteria.

When creating the master implementation, we also assessed the influence of each individual criterion on patient selection when executed against the same data source used in the original study (CCAE, Table 1). Several criteria, such as not being pregnant on the index date or being older than 17, had negligible impact on patient selection as subjects with T2D are likely to be older and, therefore, not

pregnant. The requirement of the first GLP1-RA exposure within 365 days did not have large influence on patient attrition because we initially chose the earliest event in the cohort.

Table 7. Criteria used to define master implementation and the number of subjects satisfying each individual criterion in the cohort executed against CCAE.

	Criteria	Subjects who satisfied the criteria, n (%)
Cohort entry	First glucagon-like peptide 1 receptor agonists (GLP1-RA) exposure in 2005-2017	570,664 (100%)
1	Had no GLP1-RA exposure within 365 days prior to the index date	563,245 (98.7%)
2	Had at least 365 days of prior observation	315,616 (55.3%)
3	Age > 17	569,757 (99.9%)
4	Had type 2 diabetes mellitus (T2D) within 365 days prior to the index date	430,080 (75.4%)
5	Had chronic lower respiratory disorder (CLRD) within 365 days prior to or on the index date	44,668 (7.8%)
6	Had no type 1 diabetes, cystic fibrosis, lung cancer, pulmonary embolism, pulmonary hypertension, thyroid carcinoma, conditions requiring corticosteroid therapy within 365 days prior to the index date	488,606 (85.7%)
7	Was not pregnant on the index date	565,877 (99.2%)
8	Had no insulin exposure within 365 days prior to or on the index date	402,407 (70.5%)

9	Had no dipeptidyl peptidase-4 (DPP4) inhibitor exposure within 365 days prior to or on the index date	474,365 (83.2%)
10	Had another T2D drug that started before the index date and ended on or after the index date	320,658 (56.2%)
	All criteria	6,196 (1.1%)

On the other hand, requiring a prior CLRD diagnosis, at least a year of prior observation and add-on antidiabetic therapy had a large impact on patient selection with only 7.8%, 55.3%, and 56.2% of subjects satisfying these criteria, respectively. Requiring no prior insulin exposure eliminated some subjects, but the influence of this criteria was limited by the fact that we excluded GLP1-RA and insulin combinations from the list of drugs.

Study settings

The experiment was organized as a one-day workshop, which was held as a part of OHDSI 2021 Global Symposium on September 13th, 2021. Prior to the Symposium, we invited all OHDSI collaborators to participate in the challenge. Fifty-four collaborators met all prerequisites (familiarity with the paper, OMOP CDM, Standardized Vocabularies, and OHDSI tools) and were divided into nine groups.

To ensure that the settings represent real-world phenotype development, each group had at least one informatician with extensive CDM and ATLAS knowledge and one epidemiologist or clinical expert. They were provided with access to an ATLAS instance with an empty cohort definition template. The ATLAS instance was pre-populated with the same pre-defined concept sets used in the master implementation such that the exercise was focused on the logic of the cohort definitions and not on

selection of the correct drug and diagnosis codes which in itself is challenging and would introduce significant variation. Over the day, each team separately implemented the cohort definition based on their interpretation of the paper and the supplementary materials. Groups could define any number of criteria in their implementation and apply them in any order.

Data analysis

All cohort definitions were subsequently executed against CCAE and compared to the master implementation created together with the original author. For each cohort, the number of subjects and demographic characteristics (age and sex) at index date were extracted, along with diseases and drugs used as recorded in the 365 days prior to the index date. To assess the influence of the design choices on patient selection, we calculated the agreement between each cohort created by the participants and the master cohort using the Jaccard index [266] defined as the number of subjects included in both cohorts divided by the total number of subjects in either cohort. Additionally, we extracted the variables used to describe the population in the original study and calculated the standardized difference of means between each cohort and the master implementation for each variable [267].

3.1.3 Results

Comparison of the master implementation and each team's implementations

On average, each team's interpretation fully aligned with the master implementation in four out of ten inclusion criteria; all teams had at least four criteria deviating from the master implementation. As shown in Figure 13, all nine teams fully reproduced two criteria: 1) having 365 days of prior observation; and 2) age greater than 17 years at the index date. Two additional criteria were implemented correctly by the majority of the 9 teams: 3) no conditions of exclusion within 365 days

prior (1 of 9 teams implemented this differently), and 4) no insulin exposure (4 of 9 teams implemented this differently).

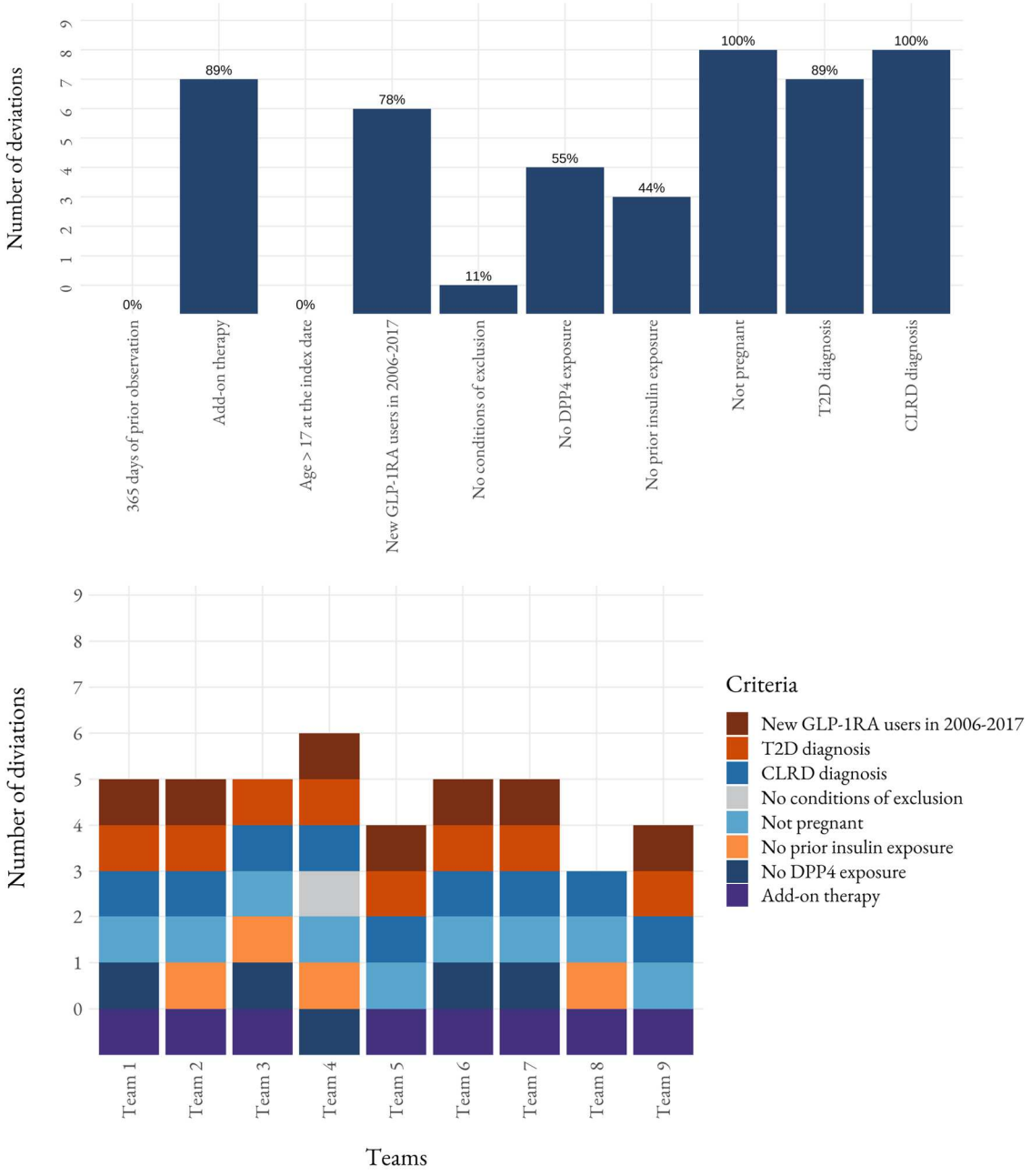


Figure 13. Number of deviations per criteria (top) and team (bottom).

Implementations of the criteria requiring complex logic were highly variable. As per the paper, the subjects had to have "… at least one inpatient or two outpatient encounters with T2D and CLRD, defined based on the presence of diagnoses or medication dispensing…during the year before index date". There were four different implementations of this criteria, which stemmed from different combinations of timing of events, their co-occurrence and combination of individual sub-criteria.

Similarly, the criterion of add-on therapy was implemented in three different ways: (a) having another antidiabetic drug on the index date, (b) having an overlapping drug exposure that starts before the index date and ends after the index date and (c) having drug exposure with a typical number of days of supply. A detailed description of the deviations per each criterion is provided in Appendix 3.1.

Influence of different choices on patient characteristics

We observed high variation in cohort size from having one third of the master implementation patient count to having ten times the cohort size (2,159 to 63,619 subjects compared to 6,196 subjects in the master implementation). Not surprisingly, the agreement between the master cohort and the teams' implementations also varied greatly (Figure 14).

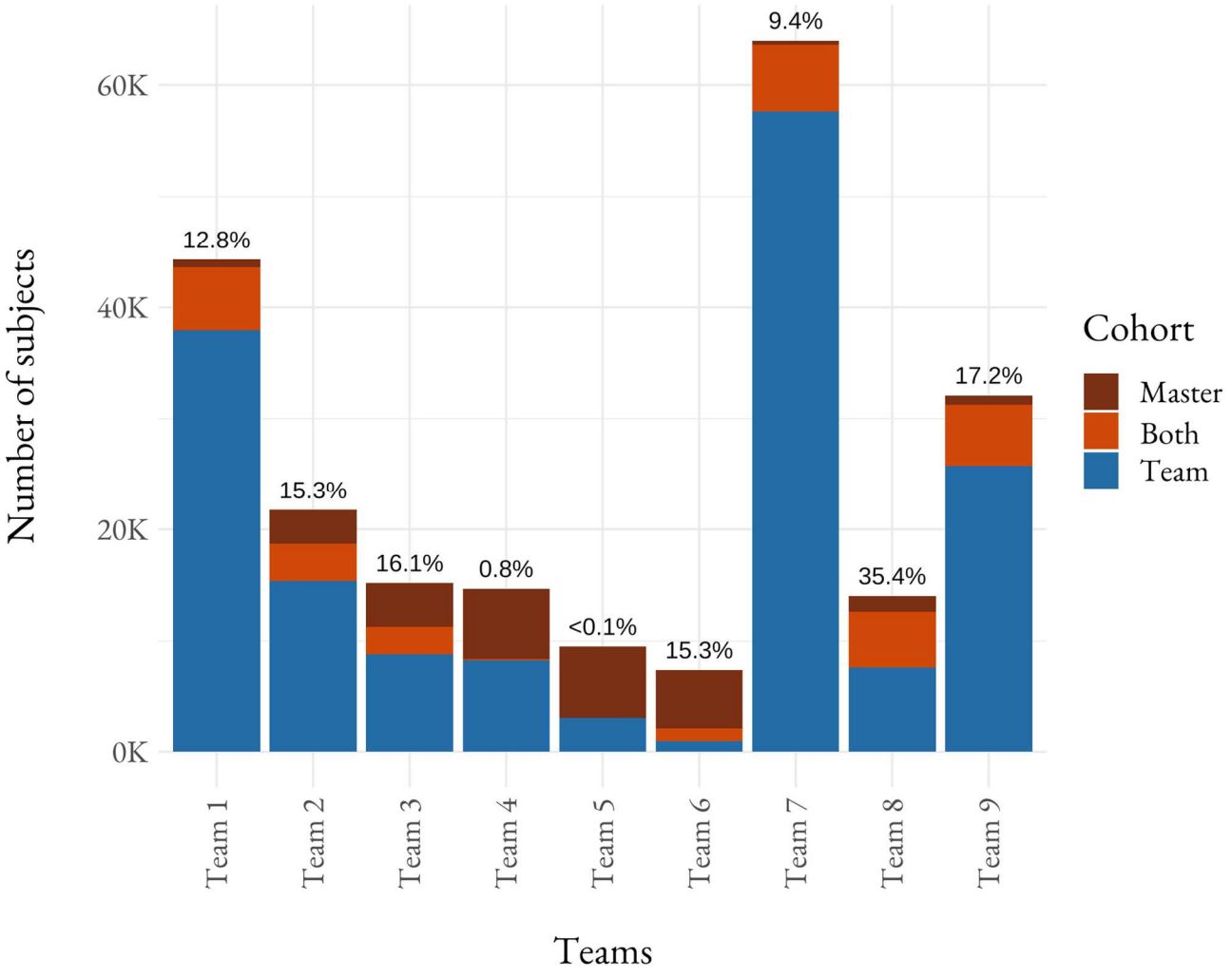


Figure 14. Cohort overlap for each team's cohort and the master implementation, number of subjects and agreement (Jaccard index, %).

Median agreement was 9.4% (interquartile range 15.3-16.2%) and ranged between 0% and 35.4%. Similarly, the teams' implementations differed from each other greatly (median agreement was 10.0% and interquartile range 0.0-17.5%).

Patient characteristics

The age distribution was similar across all cohorts with 45-64 years old being the major age group (Appendix 3.2). The gender distribution was also similar to the master implementation except for cohorts of teams 4 and 5 that had a lower proportion of females (58.3% and 57.4% compared to 66.2% in master).

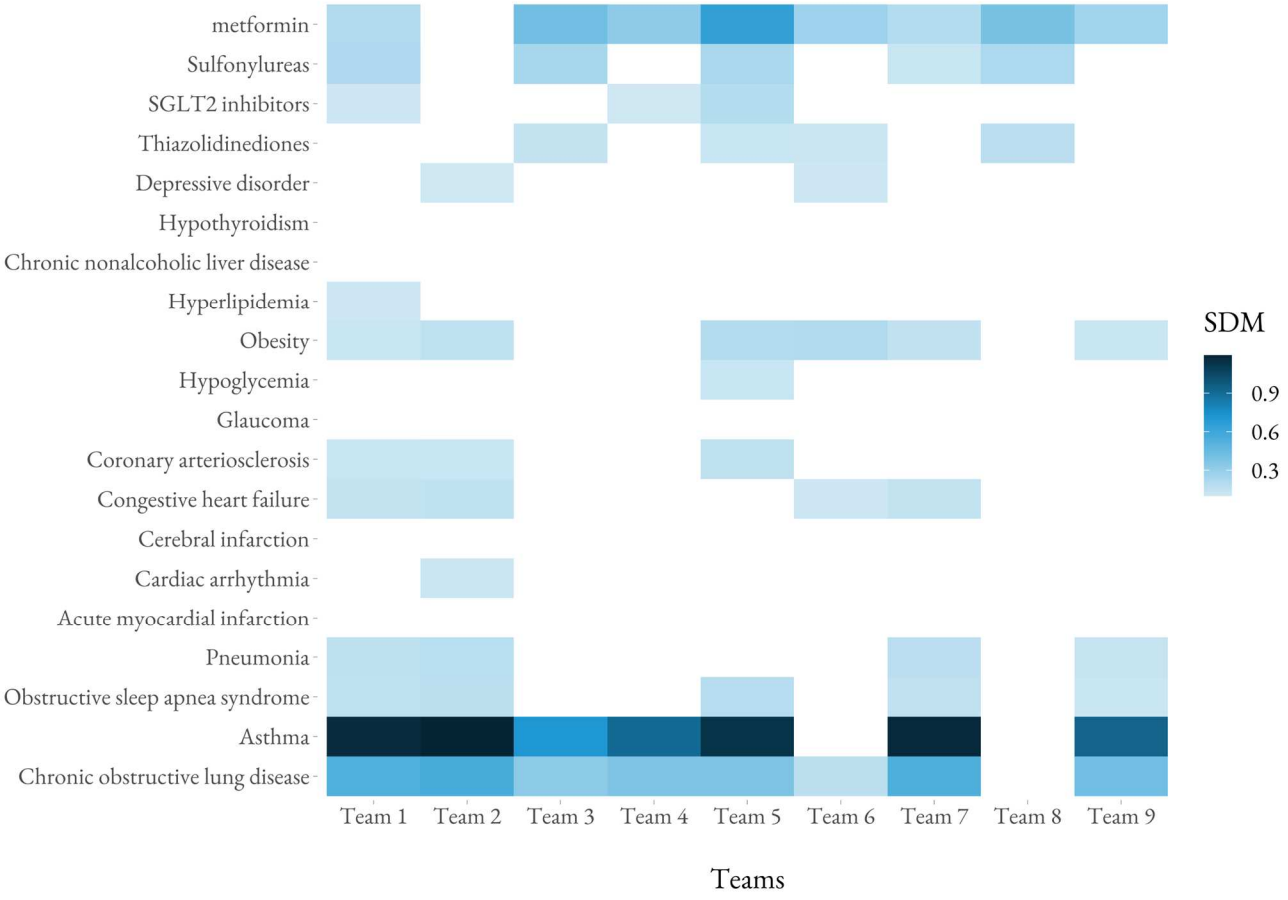


Figure 15. Difference in patient characteristics between the master implementation and teams' implementations colored based on the absolute standardized difference of means (SDM). White indicates SDM<0.1.

As shown in Figure 15, the cohort generated from each team's implementation differed from the master implementation by at least two baseline characteristics with a standardized difference of means (SDM) >0.1, and the majority of the teams differed by at least five baseline characteristics. The difference was especially prominent for chronic lung disease disorder, asthma and prior metformin exposure, which corresponded to the largest number of deviations in implementing those criteria. Cohorts were generally similar in prevalence of conditions related to T2D such as glaucoma or hypoglycemia.

3.1.4 Discussion

Here, we demonstrated that small nuances in implementation of one clinical idea produce high variability even when keeping the other elements constant. We showed that nine inter-disciplinary teams (similarly to real-world settings), given the exact same task of implementing a cohort definition based on the extensively described clinical idea using consistent development tools and pre-defined concept sets, obtained nine different cohort definitions with 52 deviations in total across a set of 10 inclusion and exclusion criteria, which resulted in vastly different patient cohorts.

Complex criteria such as type 2 diabetes mellitus and chronic obstructive lung disorder, add-on therapy and pregnancy can be defined in multiple ways and, therefore, produce especially high variation.

In fact, pregnancy can be viewed as a separate phenotype as it has a clear start and end date and can be identified based on diagnostic codes, procedure codes and laboratory measurements. Previously developed phenotypes vary from a simple diagnostic code-based [268] to a highly complex definitions [269] yet are all labeled a pregnancy.

As the extent of such variation is underappreciated in the current literature and remained ignored in phenotype development and re-use, there is limited research on methods for phenotype reporting that would reflect all of the nuances of logic implementation. An example of phenotype libraries that have the most extensive yet not standardized reporting is PheKB [166], which contains descriptive documents and figures for each phenotype as well as executable KNIME modules for some of the phenotypes.

The latter require having unified data representation for modules to be executed. More generally, understanding, reproducing and re-using phenotyping logic requires knowing the underlying data schema so the logic can be applied to the proper tables, columns, and data elements. Otherwise, the step of inferring logic must be accompanied by inferring how logic is imposed on the data elements. Having a common data model removes this inference step and directly reproduces the logic on data that have been standardized to a common format.

In this work, we were able to eliminate this source of variability by using one data instance across all implementations as well as by providing the teams with the standardized code sets and analytical tools. Nevertheless, the observed variability raises a question of validity of any clinical study that only uses one phenotype definition without exploring, examining and comparing other inclusion and exclusion criteria combinations. Similarly to reporting sensitivity analyses that examine alternative designs such as as-treated or intent-to-treat in effectiveness studies, observational studies should report the process of creating phenotype definitions including both code and logic selection. While, ideally, it should be a systematic process, more research is needed to establish it.

3.3.5 Limitations

There were limitations to the experiment. While the teams were introduced to the study before the workshop and found a full day to be sufficient to discuss and implement the definition, the activity

was limited to eight hours. We ensured that all teams had at least one clinician, bioinformatician and a team member who was familiar with the data and tools, but individual level of expertise may have varied. We selected one study as it was not feasible to have multiple teams perform multiple studies, but it is possible that the experience with this study may not be generalizable to other studies.

3.4. Portability of EHR-derived phenotypes to claims data sources[3]

As we observed in the previous aims, algorithms for identifying patients of interest need to account for data source heterogeneity stemming from disparate coding practices and populations. If such algorithms are intended to be used across a network, they are desired to achieve comparable performance on the data sources with different provenance.

Yet, the data sources differ in data cardinality and completeness. For example, administrative claims data do not generally contain the necessary information to develop accurate algorithms for disorders that require laboratory results. Phenotypes developed for claims are generally based on diagnostic and procedural codes, which may lead to decreased sensitivity or specificity.

It is unclear whether it is necessary to develop separate phenotypes for each individual data source (or a group of such) as embraced by the networks like i2b2 and Sentinel or if portable "one-size-fits-all" phenotypes can be developed to be used across the network.

In this section, I develop EHR-derived phenotyping algorithms that can be portable to claims data sources and evaluate them on four US and international data sources. I apply the insights learned in previous sections and highlight the influence of data harmonization and concept set standardization on phenotype performance and portability.

---

[3] This section is published in Journal of Biomedical Informatics. The full citation for this publication is Ostropolets A, Reich C, Ryan P, Shang N, Hripcsak G, Weng C. Adapting electronic health records-derived phenotypes to claims data: Lessons learned in using limited clinical data for phenotyping. *Journal of Biomedical Informatics* (2020)

3.4.1 Background

Phenotype portability is often limited as the algorithms reflect local patterns and practices [108]. Factors that influence portability include data fragmentation across institutions [169], difference in age [199], genetic, ethnicity and race [200] distribution across the data sources, and difference in average time of observation [270], temporal slicing of data [271] or data provenance [198].

Previous research looked at the portability of phenotypes for prostatic hyperplasia, asthma, heart failure, attention deficit hyperactivity disorder and diabetes, mainly in the eMERGE network [153,163,191,192]. Along with the factors mentioned above, these works highlight a common data model as a prerequisite for efficient phenotype development as it enables using unified tools and approaches at a cost of standardizing the data. Better portability can be achieved by harmonizing the data, using standardized terminologies and developing reproducible and standardized pipelines for phenotype development [2].

Despite some advances, the problem of phenotype portability remains largely unsolved with more in-depth research needed. Here, we examine the requirements for EHR-derived phenotypes to be portable to administrative claims data sources using chronic kidney disorder (CKD) as an example.

As CKD is a highly complex disorder that often remains undiagnosed, generating and evaluating algorithms for CKD is often a time- and labor-intensive process prone to bias. CKD phenotyping depends heavily on the origin of the data as estimation of kidney function is based on laboratory or instrumental tests. EHRs can accurately capture the decline in kidney function because is reflected in glomerular filtration rate (eGFR) and microalbuminuria level recorded in the EHR [272]. However, such detailed information is not gathered in administrative claims data, which limits their ability to support clinically accurate algorithms. Because of that limitation, an alternative approach of using diagnostic

codes to identify patients with CKD has gained popularity. Studies that focused on identifying CKD in administrative claims data have been in agreement for the low sensitivity of algorithms that use ICD-9 codes, ranging from 11% to 32.7% [273–276].

We hypothesize that codes that likely indicate the presence of the disease, i.e., codes for dialysis and kidney transplantation, procedures that are exclusively used to treat CKD, can consistently improve algorithm performance. In this study, we generate and validate a billing code-based algorithm for CKD to test our assertion that the positive predictive value of a CKD phenotype can be improved by adding other codes indirectly related to the diagnosis of chronic kidney disorder. Given limited portability of existing CKD phenotypes [277], we externally validate our algorithms on different data sources and identify the possible data discrepancies and inaccuracies that influence portability.

3.4.2 Methods

We first replicated a validated eMERGE gold standard EHR-based algorithm for CKD. Its backbone is eGFR measurement as an indicator of kidney function. We then adapted the algorithm to generate test algorithms intended for administrative claims and compared their performance to that of the gold standard.

Gold standard

The gold standard algorithm was developed and tested on the CUIMC EHR data, and further validated by chart review [278]. This algorithm follows the National Kidney Foundation's (NKF) Kidney Disease Outcomes Quality Initiative (KDOQI) CKD staging recommendations and is based on eGFR (G-stage) and proteinuria (A-stage) measurements. To calculate eGFR, we extracted the data about age, gender, race, serum creatinine measurements and used them in Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation [279].

Although the units of measure are pre-specified in the formulas, the typical units of included measures vary in some countries. For each laboratory test, we defined the possible spectrum of units of measure (e.g., gram, milligram, international unit, millimole) and created conversion tables to translate them to standard units used in the equation (Appendix 3.3). We then calculated A and G stages and excluded the patients that experienced a decline in kidney function co-occurring with acute states (acute kidney injury).

To define acute kidney injury as well as the other states used in the algorithm (kidney transplantation, dialysis, other kidney disorder), we leveraged OHDSI Standardized Vocabularies to extend the original expert-reviewed concept sets. For these concept sets to be used across disparate data sets, we leveraged 'Is-a', 'Part-of' (additional SNOMED followed by', 'has due to', 'has associated procedure' relationships) and 'Maps to' relationships stored in the OHDSI Standardized Vocabularies. Instead of limiting concepts to a vocabulary, we took the hierarchies of the vocabularies above and followed the descendant relationships to cover more granular concepts derived from the ancestors of interest. Additionally, we used the crosswalks to the non-standard vocabularies (e.g., ICD-9(CM) and ICD-10).

The algorithm produced several categories of patients: CKD Case (includes end-stage renal disease associated with transplant, end-stage renal disease associated on dialysis, CKD stage I-V, CKD Control (eGFR indicates no renal failure) and CKD Unknow/Indeterminate (patients with no eGFR or decline in kidney function associated with acute disorders). The flowchart for the gold standard can be found in Appendix 3.4 and the executable script can be found on GitHub [280].

Test algorithms

We created four billing code-based test algorithms, from simple to complex:

1. Patients with at least two occurrences of CKD diagnosis within two years.

This algorithm represents a typical approach to CKD identification in administrative claims data. This algorithm requires patients to have at least two codes from the CKD code set within two years, which is the simplest approach as it uses only CKD diagnosis codes and does not leverage any additional diagnosis, procedure or device codes.

2. Patients with at least one occurrence of CKD, dialysis or kidney transplant, and at least one additional occurrence of CKD prior to or after the index date (the date the inclusion criteria are met), excluding acute states.

This algorithm includes patients with the kidney transplantation and a diagnosis of CKD within six months before or after transplantation. The algorithm also includes occurrence of kidney dialysis and a diagnosis of CKD within six months before or after dialysis, or two occurrences of CKD within a year, and excludes patients who experienced an acute state (sepsis, shock) or acute kidney failure within 30 days prior to the index date.

3. Patients with at least one occurrence of CKD, other kidney disorders, renal impairment, dialysis or kidney transplant, with at least one occurrence of CKD prior to or after the index date, excluding acute states.

Additionally, we included renal insufficiency and other renal disorders that might indicate CKD (renal disorders in systemic disease, hypertensive renal disease, and diabetic renal disease). This algorithm may be viewed as a more complex one, but it is also may be characterized as a more sensitive one as we include additional diagnosis codes, which represent conditions that lead to CKD (renal disorders associated with hypertension, diabetes mellitus type II and systemic disorders).

4. Patients with end-stage renal disorder (ESRD).

We also created a narrower cohort that focuses on patients with CKD Stage V. We included patients with the kidney transplantation and a diagnosis of ESRD within three months before or after transplantation; occurrence of kidney dialysis and a diagnosis of ESRD within three months before or after dialysis, or two occurrences of ESRD within a year, excluding patients who experienced an acute state (sepsis, shock) or acute kidney failure within 30 days prior to the index date.

Data standardization and harmonization

To make robust phenotypes that will be portable to other data sources we performed measurement and unit harmonization.

First, we identified the possible list of units that were associated with measurements used to calculate eGFR. We created a conversion table to standardize the units that may be used with a measurement. For example, depending on the provenance of the data and local standards, creatinine can be measured in mg/dL or μmol/L. For such cases, we picked one standard unit and converted to others to it to ensure that all values have the same scale.

Second, we set an approximate threshold for each measurement to identify extremely low or high values; an example of such extreme values can be the height of 100 meters. We then eliminated those extreme values since they might have biased kidney function assessment.

Finally, we leveraged OHDSI Standardized Vocabularies to obtain a comprehensive list of concept codes for patient identification. As opposed to manual selection of codes, we used SNOMED hierarchy, which allowed us to get all relevant codes. As our network study was run on multiple datasets, it was crucial to capture clinical codes from disparate source vocabularies (such as CPT4, HCPCS, ICD-10(CM), ICD-9(CM), etc.). Instead of creating lists of possible codes for each vocabulary,

we used mappings provided by the OHDSI Standardized Vocabularies to obtain a comprehensive list of codes.

Test algorithm evaluation

For each of the algorithms we used the gold standard identified the number of true positive and false positive subjects and calculated PPV defined as true positives/(true positives + false positives) against the gold standard.

We executed gold-standard, test algorithms and performed evaluation on four EHR datasets: CUIMC EHR, AmbEMR, OncoEMR and DA Germany (Table 1). These datasets were chosen because they contained a sufficient number of patients with these diseases and represented diverse settings of clinical care (outpatient and inpatient visits as well as primary care visits and specialists' visits) and countries (US and Germany).

3.4.3 Results

The number of patients identified by the gold standard algorithm varied greatly from 171,948 patients in CUIMC EHR to 3,438,251 in AmbEMR. We approximated the prevalence of CKD using the patients generated by the gold standard and compared this number to the size of the adult population. The prevalence ranged from 1.8% in German ambulatory population to 18.9% in the US oncological population (Figure 16).

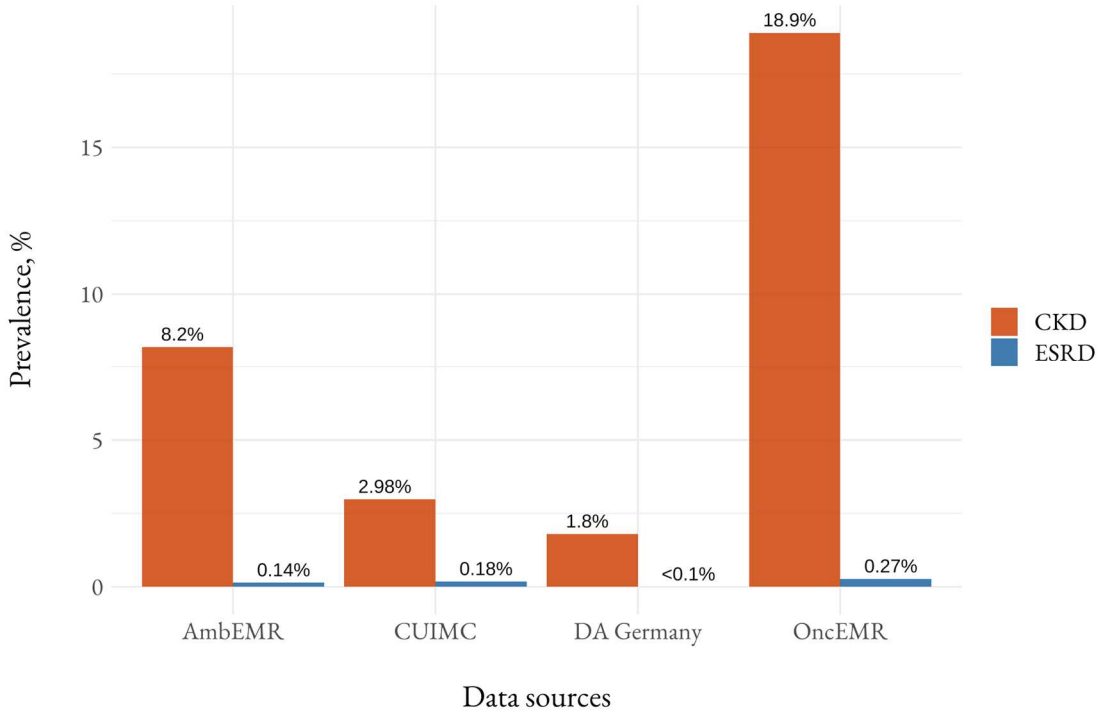


Figure 16. Prevalence of chronic kidney disorder and end-stage renal disorder, % in four datasets.

Comparisons of our test algorithms to the gold standard are provided in Table 8. Compared to the algorithm that utilizes diagnosis codes only (algorithm one), test algorithm 2 and 3 identified a significantly more patients than a typical diagnosis-based approach while preserving comparable PPV.

Algorithm two was the broadest one and included not only occurrences of CKD as an inclusion criterion, but also the generic concept of kidney impairment and other kidney disorders (kidney disorders in diabetes mellitus, systemic disorders and hypertension). This allowed the consistent performance of the algorithm across disparate datasets with different coding practices.

Table 8. Number of patients and positive predictive value (PPV) of the algorithms for CKD compared to the gold standard. T – kidney transplant, D – dialysis, SKD – suspicious kidney disorders

Dataset	Two CKD diagnosis codes		CKD or T + CKD or D + CKD		CKD or T +CKD or D +CKD or SKD + CKD	
	Patients, n	PPV, %	Patients, n	PPV, %	Patients, n	PPV, %
CUIMC EHR	45,444	45.1	64,007	43.7	70,445	41.7
AmbEMR	687,898	61.3	1,000,555	61.6	1,133,844	61.0
OncEMR	18,843	52.9	78,868	62.8	79,973	62.8
DA Germany	52,073	30.3	403,628	26.0	405,511	26.0
Average		47.4		48.5		47.9

Nevertheless, we still observed database-dependent variability in algorithm performance. For example, the performance of all algorithms on the non-US dataset had the lowest sensitivity and PPV supporting previously observed low granularity of codes in international data sources.

We also computed PPV for end stage renal disorder across all databases with the average PPV of 33.7% (51.8% in CUIMC EHR, 42.6% in OncEMR and 7.2% in AmbEMR). DA Germany dataset did not have specific ESRD codes, so our algorithm did not yield any patients on this dataset. Potentially, this algorithm can be used to detect the patients with end stage renal disorders, although it requires examination of the data to ensure that the dataset contains highly granular codes for ERSD.

3.4.4 Discussion

As phenotype development and evaluation is labor-intensive and time-consuming process, portable phenotypes can be re-used on other data sources and, in this way, facilitate rapid scientific discoveries.

While EHRs are oftentimes viewed as a common development platform for phenotyping, administrative claims data sources have been increasingly used due to large sample size and more comprehensive capture of all patient's encounters.

Here, we show that phenotypes developed and validated on EHR data can be portable to claims. We have achieved consistent performance on data sources with different country of origin and capturing different aspects of care.

It was possible with several critical components being already in place: (a) the data from all data partners was standardized to a common data model allowing us to use unified queries, (b) content harmonization through the OHDSI Standardized Vocabularies was performed, which enabled use of common code sets, (c) gold standard was already created and validated using chart review.

Even then, we had to standardize the code sets and units to account for data source heterogeneity and variable granularity, which reinforces the lessons learned in the previous sections.

Data source standardization and harmonization

Here, similarly to other studies presented in this thesis, we take advantage of a common data model.

As OHDSI OMOP CDM was created with an idea of one study fits all, it has been focusing on standardizing both the format and the content of the data. Data standardization has to be accompanied by appropriate quality assurance procedures, which would enable truly portable phenotypes.

For example, to be able to create a generalizable phenotype, one should account for the different units of measure and adjust the CKD-EPI equation appropriately. Laboratory test data might also be entered incorrectly, creating nonsense records. We observed (despite the procedures that were at place at that moment) source electronic health record data containing height that was measured in kilograms,

ratios, percent, or had negative values. Other cases are harder to identify. For example, measurement of creatinine in blood or serum has two non-overlapping normal ranges depending on the unit of measure: mg/mL or mmol/L. If the source data contain confusing or unspecified measurement units, misinterpretation of results and incorrect identification of patients with CKD happens.

To illustrate the importance of addressing these issues, we compared in Figure 17 the performance of the algorithms on the source electronic health record data and data that have been processed taking into account the issues mentioned below.

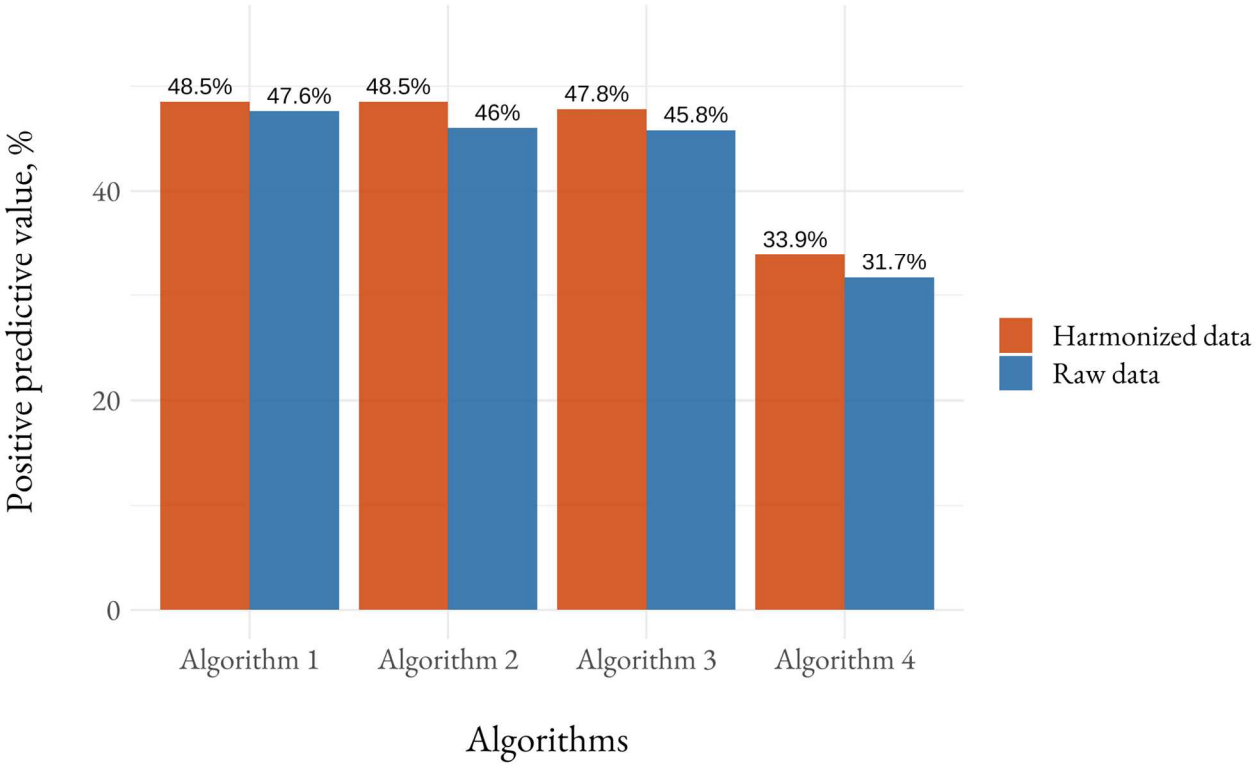


Figure 17. Overall positive predictive value of the algorithms among four databases, %.

As we can see, the positive predictive value of all algorithms improves after data standardization. Other measurement-based phenotypes regardless of the clinical area are also likely to benefit from data quality assurance rules that check the values according to a certain threshold and eliminate suspiciously

high or low values. A set of rules should be established to manipulate the data and change the units to the correct ones and drop data that cannot be interpreted.

An example the project that aims to solve these issues, which we contributed to in the early stages, is an ongoing project in the scope of the OMOP THEMIS initiative, which aims to alert the data owners when units of measure or values of measurements do not meet expected values [281]. Examples of the errors include negative values except for where appropriate or values outside of the normal range, which requires creation of manually curated reference sets.

Despite the achievements, these issues of data quality in networks are far from being ultimately solved. Not only a comprehensive process must be established, but also procedures need to be in place to ensure that all sites in the network comply to the standards.

Data source heterogeneity and granularity

Semantic standardization is required to effectively build comprehensive concept sets. An effective algorithm should either cover all possible terminologies or, as in the case of OMOP, utilize the single standard vocabulary that has relationships to the other vocabularies in a domain. Another approach to concept-set creation is pattern-driven: use string search or natural language processing techniques to obtain syntactic equivalents for concepts. This approach can only work correctly if concept names have the same syntactic patterns across different vocabularies, which is rarely true.

Here, we reinforce a need for comprehensive concept sets for portable phenotypes. To illustrate it, we extracted the codes used for CKD from the PheKB phenotype: 126 codes from SNOMED, ICD-10, and ICD-9(CM). These codes were hand-picked based on the local CUIMC coding practices, so the list does not represent the full spectrum of codes that can be used to code CKD. We also used a simple

string search with words ‘chronic kidney failure’, ‘chronic kidney disorder’, or ‘CKD’. Two hundred fifty-six concepts were retrieved. The list needed manual review or additional NLP processing, as it included negations and other modifying attributes (‘At the risk of chronic kidney disease’, ‘Chronic kidney disease resolved’). Vocabulary-driven approach using the OHDSI Standardized Vocabularies covered 278 concepts for CKD and was reviewed for accuracy by domain specialists. As we added new concepts, we were able to identify more patient records of CKD diagnosis codes with the average gain $12 \pm 2.3\%$ across test datasets.

Data source granularity has to be accounted for in a similar fashion. In our study, we found that some of the instances of chronic or acute kidney failure was partially coded as the ICD10 concept N19 “Unspecified kidney failure” of coarse granularity with no further clarification in the subsequent patient records. As chronic and acute kidney failure are mutually exclusive diagnoses, a more granular code must be inferred based on the patterns of occurrence of diagnosis, treatment pathways or other clues. Absence of more granular codes may also lead to inability to identify separate stages of CKD, where ESRD is of a particular interest. In this case, records of dialysis and kidney transplant can serve as a substitution for ESRD diagnostic codes and can be used to partially identify the patients with ESRD.

In this section we can see yet again (as in the previous section) that the specifics of implementation of a clinical idea have impact on phenotype performance and patient selection.

When implementing a clinical idea of chronic kidney failure we added kidney transplant and dialysis procedure codes to diagnostic codes as the former are used as the main treatment for CKD [282]. We also included codes for kidney transplant because this procedure is known to be associated with CKD. These algorithms identified more patients while preserving PPV, which highlights a need for exploration of different combinations of inclusion and exclusion criteria when developing a phenotype.

3.4.5 Limitations

Our work has a number of limitations. First, we performed random record review to ensure that the trajectories of the patients identified by our algorithm fit clinically relevant patterns of CKD.

Second, in the first three algorithms, we treated CKD as a single disorder without specifying a particular stage. As end stage renal disorder may be of particular interest in the clinic, we singled it out in the fourth algorithm to be able to evaluate its performance separately.

We did not compute sensitivity and specificity for our algorithms since it is difficult to distinguish between false negatives and true negatives accurately in inherently incomplete EHR data.

3.5. Knowledge-enhanced electronic profile review system

In the previous chapter, we constructed our CKD phenotypes based on the billing code portion of the EHRs and tested them against the measurement-based gold standard that was validated using a traditional chart review method. An alternative to this approach, which is using linked EHR-claims [198] or linked registry-claims datasets [283], is limited to the rare institutions that have linked data sources. Another alternative method of validation applicable to administrative claims is probabilistic evaluation [115], which is gaining increased attention but may lack broader traction due to low interpretability.

In this section we discuss the pitfalls of manual chart review for phenotype evaluation and propose a system for examining structured data elements to determine patient status (Knowledge-Enhanced Electronic Profile Review system or KEEPER). We describe its principles, utility and evaluate it on four conditions.

3.5.1 Background

Phenotyping algorithms or executable algorithms for identifying patients of interest in observational data are the backbone of observational research, including comparative effectiveness and safety surveillance, drug utilization and patient characterization studies [248]. The validity of inference highly depends on accuracy of phenotyping algorithms, which are commonly evaluated using manual chart review. This process is time- and labor-consuming, requires heavy clinician involvement, and remains the bottleneck in both data-driven and expert-based phenotyping.

Due to these limitations, it is a common practice not to evaluate newly created phenotypes but to borrow phenotypes from the literature relying on their performance metrics [284]. Nevertheless, good

performance on one data source does not guarantee portability to another, so phenotypes should be re-evaluated if used in another institution [105,153].

If evaluation is performed, the researchers typically review the charts for a small subset of patients identified by the algorithm, classify each patient as true positive or false positive and estimate positive predictive value omitting sensitivity and specificity [285,286]. Pre-clinical studies that focus on identifying the best-performing algorithms for future studies operate a larger sample size but take months [34,287–289], which is not scalable for more than one condition at a time.

As only a small sample is typically reviewed, validation results can suffer from selection bias. Previous research showed that the records of those patients who consented to supply their information differed from those who did not [290]. Phenotype-guided chart sampling strategies were proposed to mitigate this bias. They were shown to reduce variance and improve efficiency [291,292], but were developed for a specific condition (outcome-specific) and are not generalizable to other conditions.

Another challenge that undermines the validity of manual chart review is variability in execution. Previous studies reported high variability in chart abstraction and review with differences in training, high volume of information in health records and chart sparsity being the major contributors [293–295]. If patients are observed in the system regularly, the information volume grows quickly with conflicting information found in different places in the chart [296]. For example, preliminary diagnoses can propagate forward in the records with no track of the final diagnoses, discordant information can be recorded by different providers or notes can refer to earlier observations missing from the chart. On the other hand, most of the content in charts is highly redundant and useful information can be buried under duplicated notes [97].

With the advances in data extraction and mining [184,297,298], a growing body of literature uses various natural language techniques to overcome the issue of high volume by extracting the diagnosis and its severity and comparing the accuracy of extracts to manual chart review [299–304]. While these models show high flexibility and adaptability, they have to be developed separately for each condition, which limits their scalability.

Finally, chart review is simply not possible on the data sources with no charts such as administrative claims or for researchers who do not have access to identified unstructured data. While claims data offer more comprehensive patient capture as it tracks patients across all institutions, the inference from them can be perceived as inferior to EHR because traditional validation is not performed. One potential solution is using linked EHR-claims or registry-claims data sources where the former can act as a gold standard [198]. This type of validation is only available in a rather small number of institutions that have linked data sources. Alternatively, predictive models have been proposed to generate a probabilistic gold standard, use it to assign the probability of being a case to each patient identified by the algorithm and compute sensitivity, specificity, positive and negative predictive value [115]. While very promising, it lacks interpretability and subjective transparency, while reviewing charts provide an important ability to construct narratives about patients [305].

We propose that the true patient state is latent in structured data (such as ICD-10 codes) and the latter can be used to effectively ascertain patient status for phenotype evaluation. We hypothesize that three principles are crucial in this process: (a) organization of the data in the way that mimics a typical clinical diagnostic process, (b) presentation of only relevant information as opposed to the whole volume of patient structured data and (c) standardization of information extraction and representation.

We use these principles to design and evaluate a scalable chart review alternative: Knowledge-Enhanced Electronic Profile Review system (KEEPER).

3.5.2 Methods

We will describe Knowledge-Enhanced Electronic Profile Review system (KEEPER), its principles, application to four conditions of interest and evaluation.

Principles

1. Adherence to clinical reasoning

KEEPER follows general principles and steps of diagnostic clinical reasoning to patient structured data within the context of the phenotype being evaluated. For simplicity, the health outcome for which the algorithm was developed serves as a diagnosis we are clinically evaluating in a given patient.

We use elements of diagnostic reasoning to extract and present the data in several categories: clinical presentation (complaints, signs, symptoms and physical examination), history (disease history, co-morbidities, risk factors and exposures), preliminary diagnosis, subsequent diagnostic procedures, diagnoses, treatment, follow-up care and complications from structured data and present them in accordance with these conceptual elements.

2. Standardization

Both data extraction and representation are standardized across data sources and conditions. Standardized extraction is supported by a common data model (in our case, OMOP CDM) and standardized representation is tailored to the conceptual elements described above. As the steps of clinical reasoning are universal for any condition [306], the structure of data representation is unified and, as a result, disease-agnostic.

3. Dimensionality reduction

As the patient data is reviewed for the purpose of phenotype evaluation, we only extract the information that is clinically relevant to a given phenotype. We hypothesize that the structured data provides sufficient information to ascertain patient status even despite the data loss observed when going from unstructured data to structured [307].

Conceptual elements and data elements

KEEPER is built around the conceptual elements representing the typical steps clinicians follow when diagnosing a patient, which are contextualized around a disorder of interest for which the phenotyping algorithm was developed (Table 9).

The first element is clinical presentation, which consists of patient symptoms, signs, and complaints on the day they seek care (day 0 or index date). In clinical practice, physician (or healthcare team) collects current complaints, past personal and family history, assesses vital signs, performs physical examination and, based on the totality of information, makes a preliminary diagnosis.

For example, in the context of acute appendicitis phenotype, Patient X with suspected acute appendicitis (in textbook scenario) presents to the emergency room complaining of epigastric pain migrating to right lower quadrant, nausea and vomiting. Physical exam reveals fever, localized tenderness in the right lower quadrant and positive Rovsing's sign [308]. On the data level, it translates into condition codes for corresponding signs and symptoms (such as ICD-10(CM) R11.0 ‘Nausea’) and measurement codes for vital signs (such as high body temperature). As the data is already collected and the diagnostic decisions are made, we can observe condition codes for acute appendicitis or competing diagnoses such as diverticulitis or renal colic. Observing these data elements increases one’s confidence

in the diagnosis and observing symptoms typical for other conditions (such as intermittent severe pain that waxes and wanes in renal colic) or competing diagnoses decreases one's confidence.

Table 9. Conceptual elements and data representation in KEEPER.

Conceptual element	**Conceptual element in the context of the disease of interest**	**Data element**
Clinical presentation	Presence of relevant [known to be associated with the outcome] symptoms on the encounter (index date, day 0) and absence of competing symptoms	Condition codes [day 0]
Clinical plausibility	Appropriate demographics	Age, gender, race and ethnicity [day 0]
	Presence of relevant symptoms, diagnoses or treatment prior to the index date, especially recurring	Condition, drug and observation codes [before day 0]
	Presence of relevant co-morbidities and (or) pre-disposing risk factors	Condition and observation codes [before day 0]

	Absence of competing diagnoses after the index date, especially if followed by treatment	Condition, procedure, measurement and drug codes [after day 0]
Diagnostic procedures	Presence of diagnostic procedures, laboratory tests, clinical consults with other specialties, transfer to specific care sites around the index date	Procedure codes [before and after day 0]
		Measurement codes and values [before and after day 0]
		Provider and location [before and after day 0]
Treatment procedures and medications	Presence of relevant instrumental and surgical procedures performed on or after the index date	Procedure codes [after day 0]
	Presence of relevant medications prescribed or administered on or after the index date	Drug codes [after day 0]
Follow-up care and complications	Presence of relevant follow-up visits	Provider and location [after day 0]
	Presence of relevant complications after the index date	Condition codes [after day 0]

Another conceptual element is clinical plausibility, which includes specific demographics if a condition is known to be prevalent in a given group, history of disease and pre-disposing factors and conditions. Within the context of acute appendicitis phenotype, Patient X is more likely to be young [309], does not have prior recurrent abdominal symptoms, has never been diagnosed with Crohn's disease or endometriosis. If a condition of interest was chronic or had known risk factors, we would expect to observe prior episodes of care or relevant comorbidities. On contrary, observing a differential diagnosis recorded after the encounter (such as Crohn's disease), especially followed by the subsequent treatment would decrease our confidence in the diagnosis.

The next conceptual element encompasses diagnostic procedures and laboratory tests. In our clinical scenario, Patient X is sent for blood work and diagnostic imaging of the abdomen (ultrasound or computer tomography). Diagnostic findings show leukocytosis with a left shift and radiographic signs of appendicitis (enlarged appendix with wall thickening or perforated appendicitis). From the data perspective, observing these diagnostic procedures along with corresponding laboratory values would increase our confidence in the diagnosis.

Treatment procedures and medications are approached in the same way. Subsequent treatment can include a short course of antibiotics (e.g., piperacillin-tazobactam or cephalosporins in combination with metronidazole), appendectomy within a day or interval appendectomy depending on the stage and local care protocols. In our scenario, Patient X undergoes laparoscopic appendectomy and pathologic examination of the appendix reveals gangrenous appendicitis. Since the final pathologic diagnosis is consistent with acute appendicitis, the clinical case can be concluded.

With regard to structured data, pathology and operative reports are oftentimes not available unless natural language processing or manual abstraction was used to map unstructured text to structured

codes. Nevertheless, observing relevant treatment (operative or drug therapy), subsequent care and complications in structured data increases one's confidence in the diagnosis, while observing competing treatment (such as colectomy or gastrotomy) decreases it.

Presenting structured data in such a manner helps to construct narratives that facilitate patient ascertainment. For example, the first patient in Table 10 (green) is 46 year old male, admitted with abdominal pain, enlarged liver and leukocytosis. Clinical presentation is consistent with acute appendicitis or umbilical hernia, so the patient is referred to computer tomography of abdomen and is treated with a short course of antibiotic. Subsequently, the patient is diagnosed with acute gangrenous appendicitis and undergoes appendectomy. In this case, presence of relevant symptoms, diagnostic and treatment procedures as well as absence of competing diagnoses after the index date is highly suggestive of acute appendicitis.

Another example is the last patient in Table 10 (in red). 70-year-old man presented to the emergency department with symptoms suggestive of an acute abdominal problem (acute appendicitis, Barrett's esophagus and esophagitis). Given presence of hematemesis (a serious potentially life-threatening acute event with clear unambiguous presentation), we can already suspect that this was the main complaint and acute appendicitis was likely a rule-out diagnosis. Subsequent diagnostic procedures (presence of esophagogastroduodenoscopy for hematemesis and absence of computer tomography for appendicitis) and treatment (acid-reducing drugs) likely confirm that this patient did not have acute appendicitis.

Finally, the female patient in blue has the elements suggestive of appendicitis (laboratory findings and appropriate treatments) but also has the elements indicative of another condition (history of

diverticulitis and subsequent diagnosis of diverticulitis), so the choice regarding the status of such patient is left to the reviewer's discretion.

Table 10. Examples of KEEPER for three patients with suspected acute appendicitis: likely a case (green), likely a control (red) and ambiguous (blue).

Demographics and details about the visit	Presentation	Prior conditions, symptoms and treatment	Diagnostic procedures	Laboratory tests	Competing diagnoses	Treatment procedures and medications	Complications
Male, 46 yo; Visit: emergency room followed by	Abdominal pain; Acute appendicitis; Large liver; Umbilical hernia without	Abdominal pain (day -71); Abdominal pain (day -1);	Computed tomography, abdomen and pelvis; with contrast	Leukocytes (abnormal, high, day 0); Neutrophils (normal, day 0); Neutrophils/100		Appendectomy (day 23); metronidazole (3 days)	Acute gangrenous appendicitis (day 23); Acquired absence of

hospitalization (3 days)	obstruction AND without gangrene		material(s (day 0);	leukocytes (abnormal, high, day 0)			organ (day 23)
Female, 17 yo; Visit: Hospitalization (7 days)	Abdominal pain; Appendicitis; Diverticulitis of colon; Fever;	Diverticulitis of colon (day -182);	Computed tomography, abdomen; with contrast material(s); Computed tomography, pelvis; with contrast material(s) (day 5);	Leukocytes (abnormal, high, day 0/1/2/5); Leukocytes (normal, day 3/4/6/7); Neutrophils/100 leukocytes (normal, day 0/6); Neutrophils/100 leukocytes	Diverticulitis of colon (day 20);	piperacillin and tazobactam (5 days);	

				(abnormal, high, day 1-5)			
Male, 70 yo; Visit: emergency room followed by hospitalization (2 days)	Acute appendicitis; Barrett's esophagus; Esophagitis; Gastrointestinal hemorrhage; Hematemesis;	Abdominal pain (day -816); Esophagitis (day -180);	Esophagogastro-duodenoscopy, flexible, transoral; diagnostic, including collection of specimen(s) by brushing or washing, when performed (day 0);	Leukocytes (abnormal, high, day -1 and 0); Leukocytes (normal, day 1); Neutrophils (normal, day -1); Neutrophils/100 leukocytes (normal, day -1)	Diaphragmatic hernia; Barrett's esophagus; Hematemesis; Eosinophilic esophagitis; Gastrointestinal hemorrhage	pantoprazole (62 days); famotidine (2 days); ondansetron (1 days)	

Experiment

As a proof of concept, we implemented KEEPER for four conditions and conducted an experiment comparing the performance of knowledge-enhanced patient profiles and manual chart review.

We selected four conditions that represent a mix of chronic and acute conditions, rare and prevalent, those that are usually managed in inpatient and outpatient settings: acute appendicitis, diabetes mellitus type I (DMI), chronic obstructive pulmonary disorder (COPD), and end stage renal disease (ESRD).

Data extraction and gold standard

For each disease, we used eMERGE PheKB phenotypes that were developed and validated on CUIMC data [166,310–313]. All phenotypes were specified so the index date (day 0) was set to the date when the subjects satisfied all inclusion and exclusion criteria and had a diagnosis code of a corresponding condition. Once we executed phenotyping algorithms against CUIMC EHR, we selected a random subset of 20 patients for each condition and extracted relevant data elements (similar to the content of Table 10) in a semi-automated fashion.

Demographic characteristics and recorded symptoms, signs, and diagnoses on day 0 were extracted from OMOP CDM *person* and *condition_occurrence* tables respectively without any modification. Relevant co-morbidities, disease history (recorded any time before the index date), differential diagnoses and complications (any time after the index date) were extracted from *condition_occurrence* table, where selection was guided by the SNOMED-CT hierarchy and refined iteratively based on the distribution of the concepts in CUIMC EHR data. For example, for acute appendicitis we extracted all descendants of SNOMED-CT ‘Disorder of abdomen’, ‘Disorder of pelvis’

and ‘Disorder of the genitourinary system’. Risk factors such as smoking for COPD were extracted from *observation* table.

Relevant drugs (recorded any time on or after the index date) were extracted using the joint ATC-RxNorm hierarchy using grouping terms in ATC (for example, all descendants of ATC ‘Antiinfectives for systemic use’ and ‘Alimentary tract and metabolism’ for acute appendicitis) and presented on the ingredient level with days supply.

As procedures are a heterogeneous domain in the OHDSI Standardized Vocabularies, we inspected the distribution of procedure codes and manually identified relevant codes for each condition based on our clinical expertise. Measurements (laboratory tests and vitals recorded before, on and after the index date) were extracted in a similar fashion. The datasets for four conditions were then assembled similarly to Table 10 and saved as flat files. Data extraction was performed uniformly for all patients prior to their ascertainment. Chart review was performed on full patient medical records by two clinicians separately, labels for each patient were compared and iterative chart review continued until all disagreements were resolved.

Review

The experiment was conducted by four independent clinicians in two rounds (Figure 18), where two clinicians reviewed the patients with suspected acute appendicitis and patients with suspected DM1 and the other two – patients with suspected COPD and ESRD.

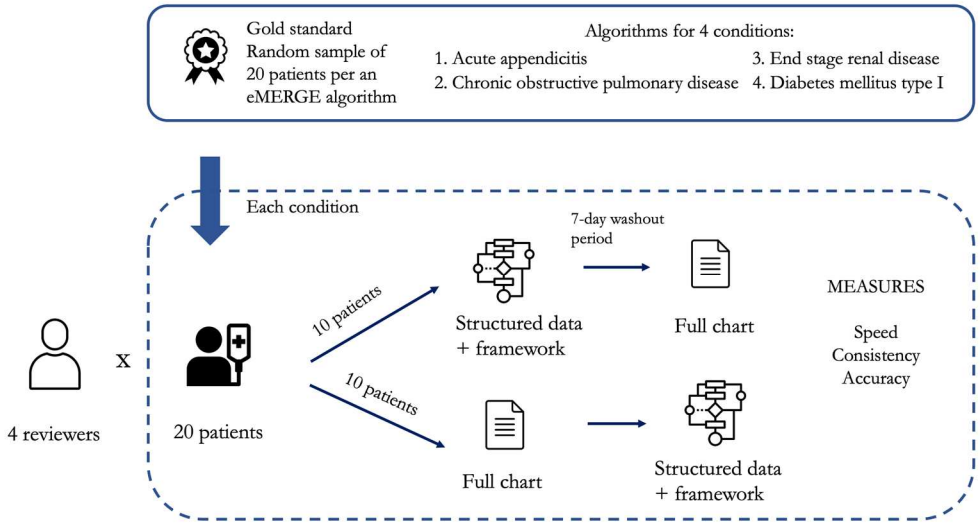


Figure 18. Overview of the proof of concept experiment design for comparing KEEPER and chart review.

We followed two-period, two-sequence crossover design, where two-period refers to two rounds and two-sequence refers to the order of studied methods [314]. For each condition, we randomly split the patients into two groups of ten, so that during the first round a clinician reviewed the profiles of patients 1-10 and charts of patients 11-20 and during the second round – profiles of patients 11-20 and charts of patients 1-10. There was a minimum of a 7-day wash-out period between rounds. Patients were assigned different identifiers to prevent carryover effect.

Each patient was classified based on the presence of the disease of interest anytime in the patient's history. Additionally, we recorded if the index date identified by the phenotyping algorithm corresponded to the date the disorder was first observed in clinical settings to assess index event misspecification.

Metrics

First, we calculated the proportion of patients classified concordantly by chart review and patient profile review (inter-method agreement) separately for each condition and overall. We used Cohen's kappa (chance-corrected agreement) to measure the agreement between patient profiles and charts for each condition as well as the overall agreement.

Second, we measured inter-rater agreement between two clinicians to assess if consistency of patient ascertainment among reviewers is improved by using standardized patient profiles. As we used fully crossed design with the goal of estimating reliability of the ratings from multiple clinicians, Fleiss's kappa was chosen as the metric for the overall agreement and Cohen's kappa for pairwise comparison [315]. The Cochran-Mantel-Haenszel test was used to compare methods across different conditions followed by Fisher exact test for pairwise comparisons [316].

Third, we compared accuracy of ascertainment against the gold standard when using full charts and patient profiles, where the accuracy was calculated as the proportion of the labels that agree with the gold standard. Proportions were compared using the Cochran-Mantel-Haenszel test.

Additionally, we compared the time spent on reviewing patient profiles and charts using the Student's t-test and performed qualitative analysis of the discrepancies in case ascertainment.

3.5.3 Results

Agreement and accuracy

We observed substantial agreement between the results of chart review and patient profile review (Table 11). Overall, 88.1% of the patients were classified similarly using full chart and KEEPER, which corresponded to Cohen's kappa of 0.71 (95% confidence interval [CI] 0.59 – 0.83).

Table 11. Comparison of chart review and KEEPER: inter-method agreement, inter-rater agreement, and accuracy.

	Inter-method		Inter-rater agreement				Accuracy	
	N (%)	Kappa (95% CI)	Chart, n (%)	Kappa (95% CI)	Profile, n (%)	Kappa (95% CI)	Chart, n (%)	Profile, n (%)
DM1	32 (80.0)	0.58 (0.34-0.82)	14 (70.0)	0.40 (<0.1-0.78)	18 (90.0)	0.77 (0.47-1.00)	34 (85.0)	35 (87.5)
Acute appendicitis	38 (95.0)	0.87 (0.69 - 1.00)	19 (95.0)	0.86 (0.56 – 1.00)	19 (95.0)	0.88 (0.64-1.00)	39 (97.5)	39 (97.5)
COPD	34 (85.0)	0.67 (0.44-0.90)	16 (80.0)	0.60 (0.28-0.92)	20 (100.0)[È]	1.00 (1.00-1.00)	34 (85.0)	32 (80.0)
ESRD	37 (92.5)	0.78 (0.54-1.00)	12 (60.0)	-0.1 (-0.3-0.1)	15 (75.0)	0.34 (-0.01-0.72	32 (80.0)	35 (87.5)
Overall	141 (88.1)	0.71 (0.59-0.83)	61 (76.3)	0.45 (0.23-0.67)*	73 (91.2)[È]	0.74 (0.52-0.96)*	139 (86.9)	141 (88.1)

* indicates Fleiss's kappa to account for two pairs of reviewers; Cohen's kappa otherwise

È indicates significant difference between two methods based on Cochran-Mantel-Haenszel test and Fisher exact test (alpha = 0.05)

Kappa ≤ 0 indicates no agreement; 0.01–0.20 - none to slight; 0.21–0.40 – fair; 0.41– 0.60 – moderate; 0.61–0.80 – substantial; and 0.81–1.00 – almost perfect agreement

Agreement varied substantially with lowest agreement between two methods for diabetes mellitus type I (moderate agreement) and highest agreement for acute appendicitis (almost perfect agreement). Even for DMI, KEEPER provided sufficient information to arrive at the same conclusions regarding patient status as with using full charts in 80% of the cases.

When comparing inter-rater agreement (agreement in patient ascertainment between two reviewers), we observed that KEEPER enabled more consistent review. Clinicians arrived at the same conclusions regarding the patients' status in 91.2% of the cases when using KEEPER compared to 76.3% when using full charts. This trend was observed for most of the conditions (diabetes mellitus type I, end stage renal disorder and chronic obstructive pulmonary disorder). In acute appendicitis, the reviewers achieved similar inter-rater agreement when using charts and using KEEPER.

We observed slightly higher accuracy of patient classification when using KEEPER compared to charts. Overall, in 88.1% and 86.9% of cases, respectively, patient classification aligned with the gold standard. In all conditions, accuracy of KEEPER was at least 80% and in three out of four conditions the accuracy was higher (albeit non-significant) or similar to the accuracy of full chart review.

Efficiency

The use of KEEPER reduced the time needed for review by more than half in both rounds. On average, chart review for 20 patients took 67 minutes (SD = 43) and patient profile review took 30 minutes (SD = 14, p-value 0.04).

There was a small, albeit non-significant difference in review time in the first round compared to the second round for both charts and profiles (charts: mean [SD] = 72.8 [45.6] first round and 61.0 [47.6] second round; profiles: 32.3 [14.0] and 28.3 minutes [16.3], respectively).

3.5.4 Discussion

In this study, we examined application of the clinical reasoning process to structured patient data for phenotype evaluation. It has long been posited that crucial information about the patient state, diagnoses and symptoms is most fully and accurately recorded in unstructured free-text notes and that only the notes can serve as the gold standard in phenotype evaluation.

Unstructured data offers great opportunity for expression, allowing clinicians to both interpret other providers' narratives and create their own [305]. As a result, multiple endeavors in natural language processing aiming at improving phenotype development and evaluation by capturing this aspect of free text were designed and implemented. These approaches remain largely disease-specific, which limits their scalability [303,317–319].

KEEPER mimics this aspect of the chart review process and can complement probabilistic methods for phenotype evaluation [115], improving transparency and trustworthiness.

Representing the data in a standardized format according to the elements of clinical reasoning enables sense-making and efficient interpretation. For example, we can construct a narrative about the last patient in Table 10 (in red). 70-year-old man presented to the emergency department with symptoms suggestive of an acute abdominal problem (acute appendicitis, Barrett's esophagus and esophagitis).

Given presence of hematemesis (a serious potentially life-threatening acute event with clear unambiguous presentation), we can already suspect that this was the main complaint and acute appendicitis was likely a rule-out diagnosis. Subsequent diagnostic procedures (presence of esophagogastroduodenoscopy for hematemesis and absence of computer tomography for appendicitis) and treatment (acid-reducing drugs) likely confirm that this patient did not have acute appendicitis. As we see in this experiment, such narratives provide similar level of accuracy compared to the narratives constructed based on the full charts.

As the goal is not a comprehensive patient evaluation but rather case adjudication in respect to one specific disorder, presenting only relevant information strengthens the narratives. Free-text notes similar to structured patient data can contain an overwhelming volume of redundant and oftentimes conflicting and inconsistent information, obstructing inference about the patient state [294]. Indeed, we observed that contradicting information in different places in the charts was a source of disagreement in patient ascertainment among reviewers. For example, some patients with bronchial obstruction (evaluated for COPD) did not have history of asthma in the recent notes, while their earlier records (sometimes going back 10 years and more) indicated a diagnosis of asthma, montelukast (a drug almost exclusively used for mild and intermittent asthma) and bronchodilator use, which undermined the reliability of the later diagnosis of COPD. Finding this information required scrutinizing tens of clinical notes, oftentimes duplicated, which lengthened reviews and decreased accuracy [320]

On contrary, KEEPER represents only relevant data in a structured way, which improves agreement between the reviewers. Similarly to our finding, standardized protocols and practices were shown to improve inter-rater reliability in the chart abstraction process [321,322].

Reviewing profiles was substantially faster and the time spent on profile review was relatively consistent across the cases and rounds. Therefore, researchers can review more patients with KEEPER, thus enabling more reliable estimation in clinical studies. It can be especially useful in patient adjudication for safety research where rare outcomes require larger sample sizes [323].

As we observed heterogeneity in performance across different conditions, more research is needed to disentangle factors influencing inference from structured data.

In COPD, the factors that contributed to lower accuracy compared to other conditions included inability to (a) easily interpret the results of pulmonary function tests to distinguish COPD from asthma or chronic bronchitis and (b) ascertain the cases when no results of pulmonary function tests were available. Similar challenges were encountered in full chart review, especially if the results of pulmonary functions tests were contradictory or inconclusive.

Disagreement in patient ascertainment among reviewers can partially be explained by differences in clinical training and expertise and different approaches to chart review. One scenario involved starting with the day 0 provided to clinicians and reviewing patient data around the day 0 first, moving sequentially along the longitudinal patient record. Another scenario involved starting at the data elements that carried the most accurate perceived information (such as pathology reports for acute appendicitis or specialty notes associated with laboratory values for the other conditions) and then retrospectively reconstructing the case.

Standardization of data representation in KEEPER partially mitigated this issue leading to higher inter-rater reliability.

KEEPER is efficient if the structured data contain the necessary elements for valid inference. While it is likely to be true for prevalent conditions and conditions requiring intensive therapy [324–

326], it is commonly acknowledged that asymptomatic conditions and some co-morbidities are underrepresented in structured data [327]. Similarly, structured data and billing codes are not likely to capture conditions associated with privacy concerns [328].

It is therefore not clear to what extent the performance observed in this study can be replicated on claims data sources, especially for those conditions whose diagnosis is heavily measurement-based. In our example, sensitivity of KEEPER may be low when attempting to classify patients with COPD or ESRD on claims data sources as there are patients who do not receive specific treatment and, therefore, can be misclassified as controls.

Future work

As we hypothesize that the information presented in a structured way may facilitate decision-making, future work includes repeating the experiment with informaticians and epidemiologists with no medical training to assess the ability of patient profiles to convey compelling stories about patient state.

Another area of future work includes designing and building an executable package and a user interface to disseminate KEEPER and connect it to a larger stack of OHDSI tools. It will allow seamless integration of phenotype development, cohort execution, cohort diagnostic and phenotype evaluation. For this solution to be scalable, relevant information has to be extracted in an automated disease-agnostic fashion. There are many works on identifying similar concepts or information, including lexical, ontological and data-driven approaches [329–333] to can be leveraged to accomplish this task.

We will likely adapt or develop a mixed-methods approach that can identify relevant but not necessarily semantically similar concept, concepts from different domains (such as laboratory tests relevant to a given disease) and clinically meaningful concept pairs (such as diagnosis-differential diagnosis pairs [334]).

3.5.5 Limitations

In this experiment, we showed the performance of KEEPER in CUIMC EHR data. As there may be high variability in the available structured data elements in other data sources, the findings may not be generalizable to the institutions with higher expected information loss from charts to structured records.

We conducted the experiment for four conditions and while these conditions represent a spectrum of disorders requiring different levels and settings of care, the results may not be generalizable to other conditions.

3.6 Chapter summary and lessons learned

In this chapter we discussed the measurement error in phenotyping in distributed data networks, which is one of the bottlenecks in timely and reliable evidence generation.

We went through each step of expert-based phenotyping starting from concept sets to cohort definitions to phenotype evaluation, highlighting a need to make data-driven decisions at each step.

As showed across all sections, the prerequisites for robust and accurate phenotyping in networks are data source standardization and content harmonization that enable standardized and scalable processes.

While there are some advances in building pipelines for seamless data processing that minimizes information loss, data harmonization and data quality assessment, more informatics solutions are needed for researchers to be able to efficiently determine data source relevance and quality. More solutions are needed comprehensive assessment of data plausibility and missingness and scalable incorporation of other data types like imaging or genetic data.

Robust and portable phenotypes also require comprehensive concept sets, which must account for data source heterogeneity and variability in granularity. Phenotypes with code sets created based on the literature or local data instance are more prone to measurement error and index event misspecification.

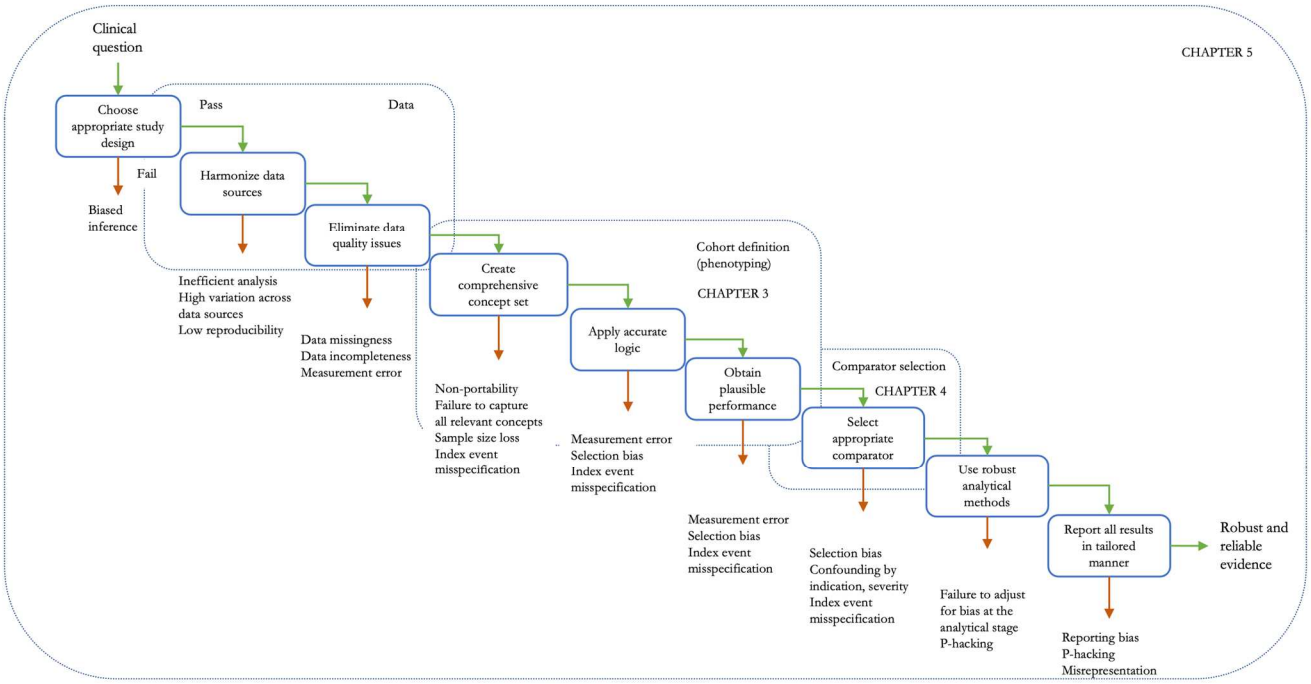
Similarly, the choice of inclusion and exclusion criteria and the order of their application influences patient selection, and, in turn, may increase measurement error or introduce bias. While we do not have a readily available solution for accounting for this variability, we can recommend explicit documentation and reporting (ideally in a structured and reproducible form) as well as examination of

different combinations of criteria and design choices. Future work can potentially focus on formalizing the set of sensitivity experiments that should be conducted to explore these combinations.

Robust phenotypes require efficient and scalable evaluation. Manual chart review is not always possible, is variable and time-consuming and probabilistic approaches are very promising yet hard to interpret. As we demonstrate, structured data can be used instead to achieve similar accuracy of evaluation, better inter-rater reliability, and substantially faster review. Potentially, the patterns we learned in our study can inform future work on meaningful examination of aggregated patient data in the cohorts to derive overall phenotype performance or estimate measurement error in an automated fashion.

Chapter 4. Addressing unexposed comparator definition as a source of bias

In the previous chapter we explored mitigating bias and reducing measurement error in one of the first steps in any retrospective observational safety or effectiveness study – identifying the cohort of patients of interest. This chapter focuses on mitigating bias in the next step – selection of an unexposed or exposed comparator cohort in cohort studies or comparison time in self-controlled case series.



Selecting an appropriate comparator is challenging. The comparator serves as a proxy for a counterfactual of the exposed population — what would have happened to those same individuals had they not been exposed — and any deviation between the comparator and that counterfactual represents a potential bias.

We partially rely on statistical approaches such as propensity score matching to mimic randomized assignment of treatment in RCTs and achieve balance in baseline characteristics between the target and comparator [335]. Nevertheless, even with appropriate statistical procedures inaccurate comparator selection may still lead to selection bias, confounding by indication or severity.

In comparative studies, a common approach is to select an alternative treatment that exhibits the same characteristics as the target treatment so that we can assume that both target and comparator treatments are administered to similar patients. Such a choice is commonly based on background clinical knowledge [68], but there is a lack of research on whether expert-based comparator selection correlates with empirical selection and if a standardized approach can be developed to improve consistency and reliability.

Having an unexposed group as the comparator as used in safety surveillance or effectiveness research presents even more challenges. The unexposed group generally represents a more heterogeneous population and does not have a clear disease onset date or exposure start date and can substantially deviate from the exposed group.

This chapter discusses the strategies for defining the unexposed comparator and is organized around drug safety surveillance (***Section 4.1***) and drug effectiveness (***Section 4.2***) tasks using COVID-19 vaccines as a case study.

First, we develop original methods for background rate estimation and interpretation to be used in observed-to-expected analysis common in drug safety surveillance. We conduct a systematic experiment on 12 data sources, examining the influence of demographics, time-at-risk and index date choices and temporal trends on background rates of 15 adverse events of special interest and provide recommendations for their interpretation in the context of design choices. The background rates

generated in this study were later used by the European Medicines Agency to assess COVID-19 vaccine safety for more than 700 million adults.

We discover high sensitivity of the estimates to the choice of database, demographic characteristics of population and time-at-risk, as well as to the choice of the index date (anchoring), and proceed by investigating the impact of anchoring on baseline patient characteristics for influenza and COVID-19 vaccines, which appears to influence baseline characteristics both short-term and long-term. We also describe original methods to empirically select an appropriate index date or event.

Finally, we assess pre-analysis bias in vaccine effectiveness research. Given mixed reports regarding the effectiveness of COVID-19 vaccines during the first two weeks after the first dose, we investigate both short-term and long-term COVID-19 vaccine effectiveness. We conduct a cohort study accompanied by the chart review to uncover selection bias, health seeking behavior-associated bias and confounding by indication and severity. Given our findings, we recommend scrutinizing the data to ensure compared groups exhibit similar health seeking behavior and are equally likely to be captured in the data and highlight the need for comparative studies when differences in baseline infection rates are present.

4.1 Assessing bias in defining an unexposed comparator for safety research

Observational healthcare data support large-scale medical product safety monitoring by detecting a possible rise in the incidence of adverse events following exposure. Drug and vaccine surveillance uses a variety of methods from observed-to-expected analysis [336] to traditional case-control, cohort or self-controlled case series studies [337], where comparator is typically an unexposed group or unexposed person time.

For example, in observed-to-expected analysis, rate of adverse events following exposure is compared with the background incidence of events occurring naturally in the population not exposed to a drug of interest [338]. While there is a substantial body of research on accurate capture of events in the exposed population, there is lack of systematic approaches to unexposed comparator definition [339–342]. As the unexposed group represents a general population and, therefore, does not have clear inclusion criteria or index date, there is a need for systematic evaluation of the parameters of the unexposed group and how such parameters influence background rates.

In ***Section 4.1.1*** we conduct a systematic experiment on 12 data sources, examining the influence of demographics, time-at-risk choices and temporal trends on background rates of 15 adverse events of special interest and provide recommendations for background rates calculation and interpretation. We then proceed (***Section 4.1.2)*** with investigating the influence of index date choice on baseline patient characteristics for influenza and COVID-19 vaccines.

4.1.1 Sensitivity of background rates to the choice of unexposed comparator [4]

4.1.1.1 Background

There is a body of research on accurately inferring observed rates of adverse events from spontaneous adverse event reporting systems such as FAERS, VigiBase or EudraVigilance [343–346] or patient portals and social media such as DailyStrength, PatientsLikeMe and Twitter [347,348].

Meanwhile, there is less research on accurate and consistent capture of baseline incidence rates (IRs), which becomes especially challenging for safety monitoring in new patient populations or mass preventative measures such as vaccination campaigns [338,349]. For example, the protocols for background rates for COVID-19 vaccine safety published by several regulatory bodies differ in data sources used, requirements for prior observation periods, index date and outcome definitions [350,351]. Recent papers on estimating background rates of adverse events of interest for COVID-19 vaccine also used heterogeneous definitions and settings [352–354]. Such discrepancies may result in producing different incidence rates and obscure their interpretation.

Lack of a common framework and high variability in approaches for baseline IR calculation results in high heterogeneity of IRs in most of the meta-analyses of IRs [355–358]. As baseline IRs directly influence the study estimates and conclusions we draw, the current lack of guidance on baseline rate calculations may lead to biased inference about vaccine or drug safety.

---

[4] This section is published in Frontiers in Pharmacology. The full citation for this publication is Ostropolets A, Li X, Makadia R, Rao G, Rijnbeek PR, Duarte-Salles T, et al. Factors Influencing Background Incidence Rate Calculation: Systematic Empirical Evaluation Across an International Network of Observational Databases. *Frontiers in Pharmacology* (2022)

Several factors have been noted to influence baseline rates, such as age [359–362], sex [362–365], race [366–368], patient location [359,365,368,369] and primary healthcare institution [360,362,370,371]. For example, the studies reported up to a 10-fold difference in IRs of adverse events in different age groups [338], up to a 20-fold difference in IRs across different data sources [362]. Nevertheless, these factors have not been studied systematically.

The impact of time-at-risk (TAR) start and duration has not been systematically examined either. In the exposed group TAR choice usually based on the pharmacokinetics and pharmacodynamics of the drug. On contrary, in background rates TAR is often set to a year and it is not clear if this discrepancy impacts study estimates.

Another gap in research is related to the starting point (time zero or index date) used to estimate baseline IRs. Most of the studies use an arbitrary calendar date for time-at-risk start, which can be the date patients satisfy the inclusion criteria or start of the year for annual IRs. On the other hand, anchoring (i.e., indexing) time-at-risk intervals on a healthcare encounter may be more appropriate but is likely to be associated with observing more adverse events due to the impact of administered drugs or detection bias.

As there is insufficient empirical study of factors influencing incidence estimation and the magnitude of such influence, we systematically analyze the parameters influencing background rate estimation and discuss implications for interpreting incidence rates using the incidence rates for adverse events of special interest for COVID-19 vaccines as an example.

4.1.1.2 Methods

Study design

Our primary research question was: "How does the selection of analysis parameter choices (such as target population, anchoring event, time-at-risk, and data source) influence baseline incidence rate estimation?" To address it, we identified the choices related to each part of the incidence rate estimation (Figure 19) and specified a set experiments to estimate the sensitivity to those parameter choices.

All parameter choices are described separately below followed by the description of the experiment ("Sensitivity experiment").

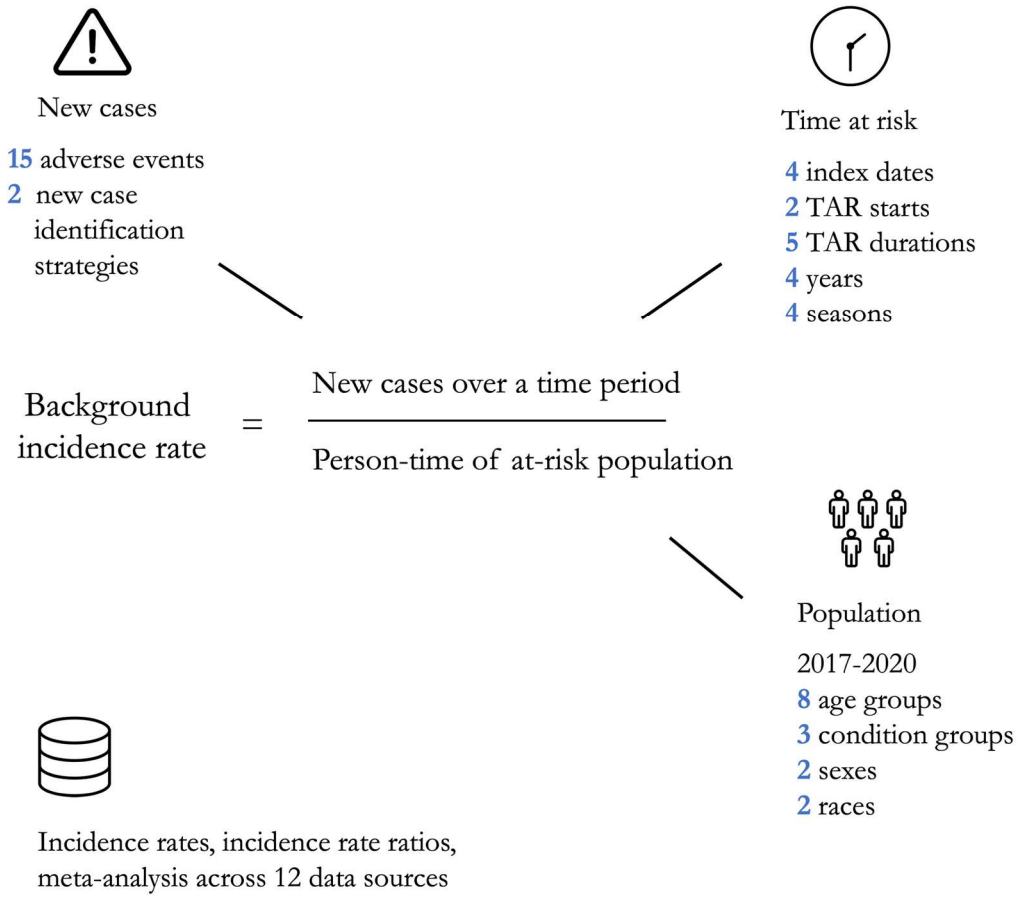


Figure 19. Baseline incidence rate calculation and its elements in the context of study of sensitivity of background rates.

Phenotype development

We used 15 outcomes outlined in the "Background Rates of Adverse Events of Special Interest for COVID-19 Vaccine Safety Monitoring" protocol published by Food and Drug Administration Center for Biologics Evaluation and Research [350].

We followed OHDSI phenotype development and evaluation pipeline described in Chapter 2 to translate and expand the phenotype definitions from the above-mentioned protocol to ensure that the clinical codes cover US and non-US data sources. This was done through translating the codes to the standard representation in the OHDSI Standardized Vocabularies and iteratively expanding the code sets using PHOEBE [372]. We also systematically examined each cohort to assess patients' characteristics such as demographics, baseline co-morbidities, drug use, procedures and health utilization using CohortDiagnostics [373].

We did not examine phenotypes requiring an inpatient encounter on the outpatient data sources (Australia EMR, DA Germany, DA France, ICPI Netherlands). We also excluded the phenotypes that did not yield patients on given data sources, as well as age strata less than 55 years for MDCR. Results for transverse myelitis in JMDC and narcolepsy in Optum EHR were removed due to failed cohort diagnostics.

Target population

The base population was the patients observed in the database at any time during 2017-2019 with at least 365 days of prior observation. The background rates calculated in this base population were published in a standalone clinical manuscript to guide decision-making for COVID-19 vaccines [16].

We also selected several subgroups of interest for COVID-19 vaccine based on health state and behavior (Figure 19). For patients with a well visit, the latter was defined as a healthcare encounter associated with CPT4 codes representing well visits. A chronic condition visit was defined as a healthcare encounter with at least one condition diagnostic code associated with a higher risk of complications as defined by CDC. Pregnancy episodes were constructed using a published algorithm [269]. The populations were further stratified on age (0–5, 6–17, 18–35, 36–55, 56–64, 65–74, 74–85, > 85), sex (male, female) and race (White, Black). Race was extracted from the patients' electronic health record (CUIMC EHR and Optum EHR) or commercial claims (Optum SES) for whom a race field was populated.

Time-at-risk

We anchored the time-at-risk on a random date, health care visit, well visit or influenza vaccination, and we applied several time-at-risk interval durations (Figure 19). We studied years 2017, 2018, 2019 and 2020 separately, and we studied seasonal intervals as dates 1/1–3/31, 4/1–6/30, 7/1–9/30 and 10/1–12/31 in each year. We also compared the COVID-19 pandemic (4/1/2020–9/31/2020), to the same period in 2019.

Sensitivity experiment

We performed calculations for each combination of outcome, target population and time-at-risk. We calculated incidence rate as the ratio of the number of cases to the total person-time the population was at risk (from cohort start date to the end of time-at-risk period, occurrence of an outcome or loss to follow-up whichever comes first).

To make comparisons between the incidence rates observed under different analysis settings, incidence rate ratios (IRR) were computed, holding all parameters constant except for the target

parameter of interest. Comparisons using IRR included: male versus female patients, White versus Black patients, no 'at risk' comorbid condition versus >=1 'at risk' comorbid condition, outcome-specific clean window (minimum time between outcome occurrences to be considered separate events) versus no prior events as well as comparisons of different years and seasons. For all incidence rate ratios, we conducted random-effects model meta-analyses to generate age-adjusted and unadjusted pooled IRRs and 95% confidence intervals across data sources using R package metafor version 2.4 [374]. Heterogeneity was assessed using the I^2 index [375].

Data sources

We conducted the experiment on 12 data sources (Table 1), including sources with different data source provenance (administrative claims data, electronic health record data), origin (the US, Australia, Germany, France, Japan, the UK), and representing different populations (privately insured employed patients in CCAE or patients with limited income in MDCD).

Detailed descriptions of analysis parameters for each experiment can be found on GitHub [376].

4.1.1.3 Results

The number of included patients varied from 252,212 in Australia EMR to 40,955,085 in OPTUM EHR with the proportion of female patients from 45.0% in JMDC to 59.5% in CUIMC (Appendix 4.1). The data sources covered all age groups except for patients over 75 in CCAE and patients under 65 in MDCR with patients aged 35-54 years being the most common group.

As expected, the incidence rates of the outcomes displayed a very wide range. When calculated for all age groups, target populations and anchoring events, IRs of outcomes showed more than 100,000-fold differences.

Patient characteristics

Age was the main contributor to the heterogeneity, with rates varying by up to a factor of 1,000 across age groups within one database (Figure 20). The effect of age was observed consistently across all data sources and outcomes, which highlights the extreme sensitivity of the incidence rate estimation to the age distribution of the measured population.

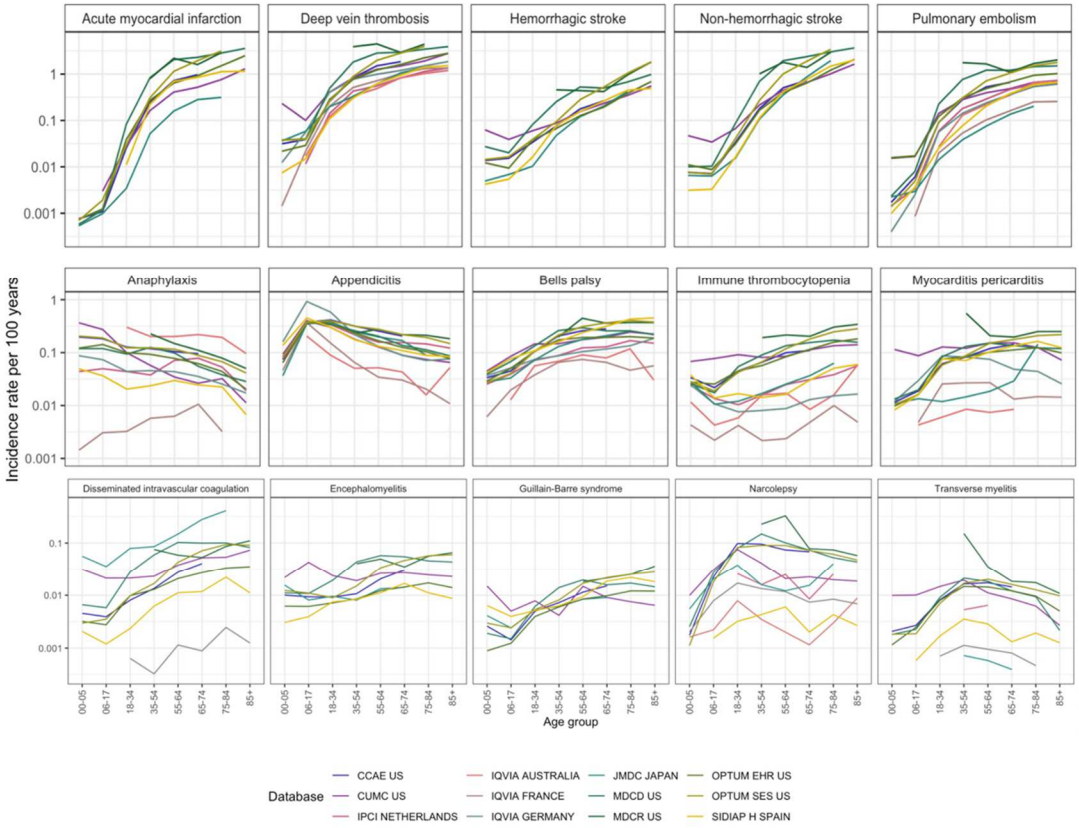


Figure 20. Incidence rates in age groups in 2017 – 2019 in patients entering on January 1 with a 365-day time-at-risk and 365 days of pre-entry observation period. Outcomes were arranged by maximum incidence per age stratum from the most common to the least common.

For sex, the IRR of incidence rates in males compared to females ranged from 0.76 to 2.17 and was statistically significant in 10 of 15 (Appendix 4.2 and 4.3). The direction generally matched the

literature: transverse myelitis was more common in females, cardiovascular conditions and appendicitis were more common in males. For most of the conditions, race did not have a substantial effect on incidence rates (Appendix 4.2 and 4.4, range 0.67 to 1.49). Disseminated intravascular coagulation, myocarditis, non-hemorrhagic stroke and pulmonary embolism were diagnosed more often in Black patients and appendicitis and Guillain-Barre syndrome were diagnosed more often in White patients.

Differences of a factor of 10 across databases were common, especially for rare disorders like disseminated intravascular coagulation or transverse myelitis. Generally, these disorders had higher incidence in the non-US data sources compared to the US data sources. Notably, disseminated intravascular coagulation had a higher incidence in Japan. All age-sex population strata showed at least 40% heterogeneity by I^2 in strata- and outcome-specific meta-analyses.

Patients with chronic conditions had significantly higher rates of all outcomes when compared to the group of patients with no chronic conditions (pooled IRR 2.16, 95% CI 1.91 - 2.44). Prior influenza vaccination was also associated with higher incidence compared to the general population (pooled IRR 1.41, 95% CI 1.30 - 1.54, Appendix 4.5 and 4.6).

Time-at-risk

When adjusted for age, anchoring was the second-largest effect, where anchoring on a visit versus anchoring on January 1st for a short time-at-risk (2 days) was associated with up to a 100-fold increase in incidence (pooled IRR 26.8 (95% CI 21.9-32.8)). The effect was attenuated for longer times at risk (Figure 21): for example, IRR for 1-28 days was 1.4 (95% CI 1.3-1.5, Appendix 4.7).

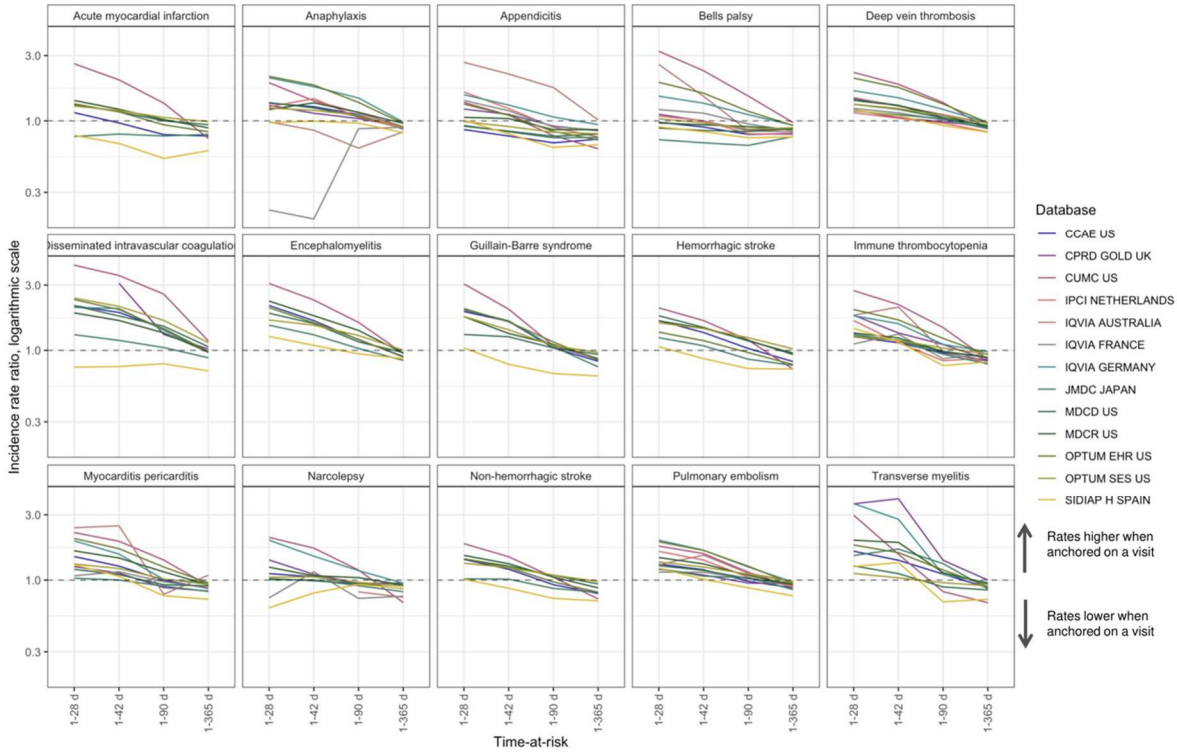


Figure 21. Comparison of anchoring on a random visit versus anchoring on January 1st in patients with a visit in the next year for time-at-risk 1-28 days, 1-42 days, 1-90 days and 1-365 days, incidence rate ratio.

Additionally, we found that when anchoring on a visit, the incidence rates for a 1-365 day time-at-risk were lower than in the group of patients with a visit in the next year anchored on January 1st. This may be explained by the fact that anchoring on a visit excluded the day of the visit from time-at-risk, while time-at-risk for anchoring on January 1st included the days of subsequent visits. Including day 0 in time-at-risk mitigated this difference (Appendix 4.7).

We observed similar trends for anchoring on a well visit or an influenza vaccination with the pooled IRR 1.21 (95% CI 1.11-1.31) and 1.17 (95% CI 1.11-1.22) respectively (Appendix 4.8 - 4.11).

Notably, incidence of Guillain-Barre syndrome was significantly increased when anchoring on an influenza vaccination and was less influenced by anchoring on a well visit or a random visit.

Time-at-risk duration influenced incidence only when we anchored on an event. When anchoring on January 1st, comparing the time-at-risk for 1 day versus 365 days showed consistently little effect across all outcomes with the pooled IRR across databases and outcomes of 1.0 (95% CI 0.93-1.08).

We observed seasonal trends for anaphylaxis, appendicitis, acute myocardial infarction, strokes and Guillain-Barre syndrome (Appendix 4.12 and 4.13). We also found a decrease in IRs in some of the data sources in 2020 compared to 2019-2017 (Appendix 4.14 and 4.15).

Incident cases

In this study, we defined incident cases as those that occurred for the first time in a given window. An alternative approach – using all patient history to identify incident cases – produced consistently smaller incidence rates for all outcomes with the pooled IRR of 0.83 (95% CI 0.79-0.87). Notably, IRRs for narcolepsy and Guillain-Barre syndrome were significantly smaller (IRR 0.69 (95% CI 0.65-0.74) and IRR 0.59 (95% CI 0.48-0.71) respectively, Appendix 4.16 and 4.17).

This observation was supported by modestly lower incidence when requiring patients to have prior observation (pooled IRR 0.94 (95% CI 0.9 – 0.99)). While this trend was not observed for all outcomes, narcolepsy, Guillain-Barre syndrome and myocarditis again were greatly impacted.

4.1.1.4 Discussion

Accurate estimation of background rates is essential for their use in safety or effectiveness studies. Background incidence rates are highly sensitive to demographic characteristics of population, so estimation requires age, sex, and potentially other adjustments to reduce bias. Such adjustments would

best be performed within the same database, but even after that, incidence rates are highly influenced by the choice of the time-at-risk start date or event.

When comparing background rates to estimated incidence rates, one must examine if the choice of anchoring is compatible between groups. If anchored, short time-at-risk intervals are associated with higher incidence, so the choice of time-at-risk requires thoughtful analysis. Similarly, the choice of clean window for defining incidence cases results in different incident rates. Finally, the choice of year and season may influence rates, albeit the influence is not prominent compared to the other factors. As opposed to demographic characteristics, which can be adjusted for in the analysis, these choices must be accurately assessed during the study design stage to mitigate potential bias. We will discuss each of the factors in more detail.

Population at risk: age, sex, race

Age and sex are the key characteristics previously shown to influence IRs [16,359,361,362,377–381]. Our study systematically explores them and shows the extreme size of the age effect in all outcomes and data sources. Therefore, one must perform age and sex adjustment when comparing background and observed rates.

Database effects

The large effect of data source choice is likely a combination of actual population differences—age, sex, race, acuity, differences in genetics and environmental exposure—as well as differences in measurement, such as collection via administrative claims versus electronic health records. Some data sources may be appropriate only for certain conditions due to their population characteristics. For example, MDCR contains patients over 65 years old, which makes it a poor choice for studying pediatric conditions. Data sources that reflect only some aspects of care (such as outpatient data sources like

Australia EMR or DA Germany) may yield different rates for conditions that commonly require hospitalization. The differences suggest that, where possible, background rates should be calculated in the database where the surveillance will be done. Where this is not possible, a broad range of databases should be used and, based on a random-effects meta-analysis, prediction intervals should be calculated for the incidence rates.

Large effect of anchoring on health encounters

Anchoring was the second most important parameter to be accounted for, at least at the shortest time-at-risk. Its influence was not quantified before and, surprisingly, was present for both random and well visits.

When studying background incidence in the context of COVID-19 vaccination (in cohort or self-controlled studies), estimation of IRs of events following vaccination is anchored on the date of vaccination. To appropriately compare it to the background rates, one has to make an assumption of the type of encounter that represents the vaccination best. For example, in a wide population that receives the vaccine based on availability, a random date may be a good approximation for the date of vaccination. On the other hand, vaccination date in patients receiving vaccine upon hospital discharge or in nursing homes may represent a strong anchor with the effect like or even greater than anchoring on a random visit. This is especially relevant for outcomes like anaphylaxis with short times-at-risk.

Influenza vaccination may serve as another proxy for COVID-19 vaccination, in terms of defining an anchoring event. But the population that receives an influenza vaccine in healthcare institutions may be different from those who receive it in pharmacies [382]. It may explain why we observed higher incidence of conditions in patients with prior influenza vaccine as vaccination in this case may be indicative of co-morbid conditions.

Muted seasonal effect and small annual increase

While previous research emphasized the influence of season on IRs [383], we observed that seasons had a minor effect on incidence. The direction of difference we observed generally matched the literature [359,384–386]. Temporal trends were moderate: incidence rates slightly increase from 2017 to 2019, which may correspond to better diagnosis or changes in coding practices. That agrees with the findings in the literature for encephalomyelitis, hemorrhagic stroke, anaphylaxis, narcolepsy, Bell's palsy [362,386–390].

Incident cases

The strategy for selecting a clean window (minimum time between outcomes) consistently influences background rates. Lower incidence of chronic conditions or conditions that are likely to occur once (such as appendicitis) when using narrow clean windows reflects potential index event misclassification. It is possible that such patients are captured later in the course of the disease, which requires thoughtful examination of the patient history to determine the true condition start date.

Using a requirement of prior observation ensures that patients were actively observed in the data source. In this study, we found that such a requirement did not produce a difference in IRs when compared to the broad population. On the other hand, it potentially reduces index event misclassification as more information about the patient is captured.

4.1.1.5 Limitations

Due to observational nature of the study, the data sources may not have complete capture of patient conditions. As the goal of the study was not to establish causality but to estimate sensitivity of incidence rates, phenotype measurement error or partial data capture should not influence the results of the study. As race is available only in three US data sources, our findings regarding race influence may

not be generalizable to other data sources or populations. Differences in incidence of adverse events of interest in different races may be attributable to differences in healthcare utilization, clinical presentation and health state awareness rather than a true difference in incidence.

4.1.2 Sensitivity of baseline patient characteristics to the choice of unexposed comparator [5]

4.1.2.1 Background

As we observed in the study of background rates of adverse events, patients indexed on an arbitrary date had lower incidence of adverse events compared to the same patients indexed on a visit, which highlights the importance of index date selection or, as we termed it, anchoring.

In this section, we evaluate two alternative selection procedures for the index date in the unexposed group based on how vaccines are administered - coupled or decoupled to another healthcare encounter. We compare these approaches for influenza and COVID-19 vaccines and investigate how anchoring influences baseline patient characteristics.

We chose influenza vaccine since under the time pressure, as observed in COVID-19 pandemic, we rely on the existing scarce body of research to assess vaccine safety and effectiveness, with influenza being an important example [18,391,392].

---

[5] This section is published in JMIR Public Health Surveillance. The full citation for this publication is: Ostropolets A, Ryan PB, Schuemie MJ, Hripcsak G. Characterizing Anchoring Bias in Vaccine Comparator Selection Due to Health Care Utilization With COVID-19 and Influenza: Observational Cohort Study. *JMIR Public Health Surveillance* (2022)

COVID-19 vaccination, nonetheless, differs in scope of vaccination, target vaccination groups and spectrum of delivery settings. The target vaccination group has shifted from the elderly and those with comorbidities in the early phases of vaccination to everyone including the healthy children [393], with some nations already vaccinating the majority of their populations [394]. COVID-19 vaccines are delivered in a wide variety of settings, from pop-up centers unconnected to health care delivery to inpatient facilities on hospital discharge. Other vaccines like those for influenza have a different delivery. They are often administered to specific vulnerable populations such as pregnant women, patients at high risk of complications or children, and are often administered during health care visits [395–397]. Therefore, the unique properties of COVID-19 vaccination may require adjusting study designs previously used for influenza vaccination, specifically the comparator definition.

COVID-19 vaccines are delivered in a wide variety of settings, from pop-up centers unconnected to health care delivery to inpatient facilities on hospital discharge. Other vaccines like those for influenza have a different delivery. They are often administered to specific vulnerable populations such as pregnant women, patients at high risk of complications or children, and are often administered during health care visits [395–397].

Therefore, the unique properties of COVID-19 vaccination may require adjusting study designs previously used for influenza vaccination, specifically the comparator definition.

3.1.2.2 Methods

Study design

We studied two types of vaccination: (a) influenza vaccine administered during 2017 – 2018 and (b) COVID-19 vaccine administered during 2020 – 2021. List of the codes used in the study can be found on GitHub [398].

For each vaccine, we mimicked two study designs. The first design (Figure 22, A) corresponds to a cohort method, where the target group was vaccinated patients, and the comparator group was unvaccinated patients. The index date for the target group was the date of vaccination; for the comparator it was (a) a date selected from the unvaccinated patient's history (not necessarily with any medical event) such that it matched the index date of one of the target group participants or (b) a visit matched to the index date of one of the target group participants. Patients in each target and comparator pair were matched on age and sex.

The second design (Figure 22, B) corresponds to a self-controlled design (case-crossover design) [12] where the cases were the vaccinated patients indexed (or "anchored") on the day of vaccination and the controls were the same patients indexed on an arbitrary date or a visit within 180-450 days prior to the vaccination date.

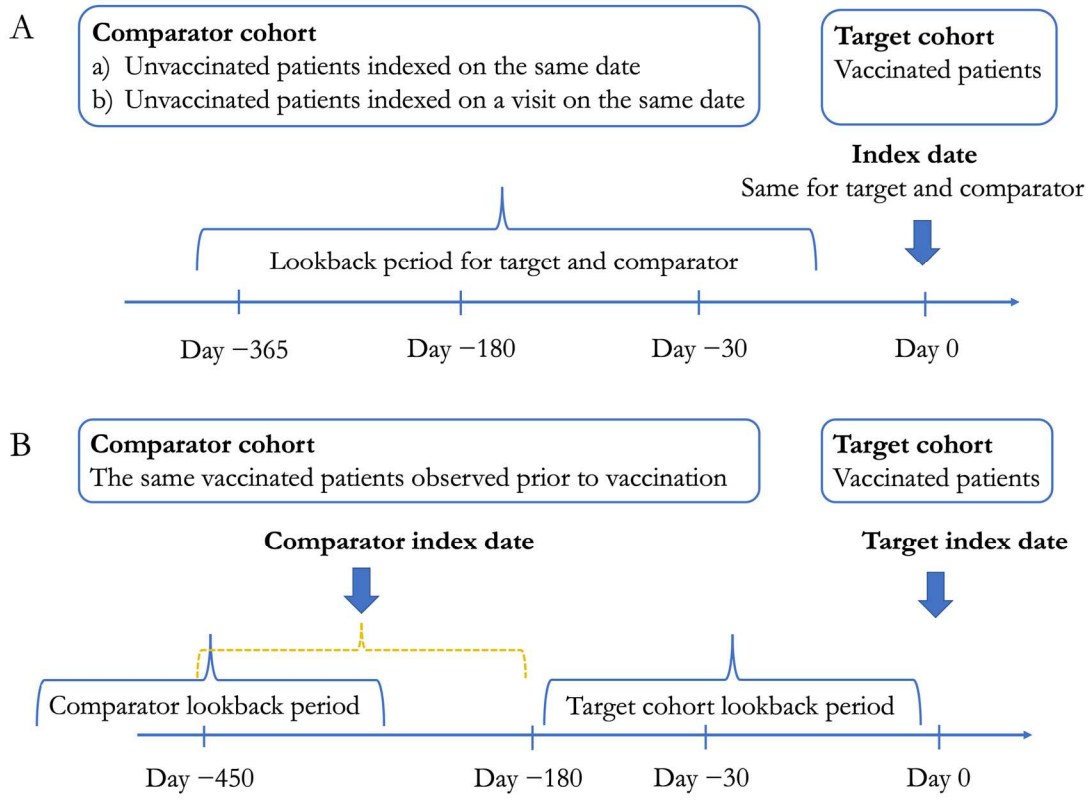


Figure 22. Overview of the study design for investigating the influence of anchoring on baseline patient characteristics.

For each group, we extracted patient baseline characteristics (covariates) recorded within five time intervals: at the index date (day 0), at the day before the index date (day -1), from 30 to 1 day prior to the index date (short term baseline covariates), from 180 to 31 day prior to the index date (medium term baseline covariates) and from 450 to 181 day prior to the index date (long term baseline covariates). Baseline covariates included all condition, procedure, measurement (laboratory tests and vital signs) and drug group codes available in patients' structured data within a specified time interval. For each covariate, we calculated covariate proportion which is the proportion of patients with a covariate recorded in their electronic health record within a given time interval along with its standard

deviation (SD) for binary variables or an average number with SD for continuous variables (such as number of visits).

We then compared the covariates in each target-comparator pair and calculated the standardized difference of means. The covariates were said to be balanced if the standardized difference of means was less than 0.1 [13-14]. The standardized difference of means for each covariate was then plotted for every time interval and target-comparator pair.

Data sources

We conducted the analysis on two electronic health record data sources: CUIMC EHR and Optum EHR (Table 1). Data source choice differed from the previous section and were selected based on the availability of both vaccines' data and captured inpatient and ambulatory aspects of care.

3.1.2.3 Results

Study population

Initial study population included 210,263 and 57,000 patients vaccinated with any COVID-19 in 2020-2021 and 60,142 and 4,991,051 patients vaccinated with influenza vaccine in 2017-2018 in CUIMC and Optum EHR respectively. Proportion of female patients was 62.7% (131,922) and 72.3% (41,204) for COVID-19 vaccinated patients and 61.4% (36,917) and 58.2% (2,906,757) for influenza vaccinated patients. Median age was 57 (Interquartile range, IQR = 39-71) and 45 years (IQR = 34-56) for COVID-19 vaccinated patients and 35 (IQR = 12-63) and 50 years (IQR = 22-66) for influenza vaccinated patients. We then matched each vaccinated population to unvaccinated population on the date, age and gender so that the distribution of age and gender between each target and comparator group was the same.

Table 12 presents the proportion of covariates with the standardized difference of means >0.1 for each comparison reflecting on the magnitude of the difference in baseline characteristics between a target and comparator.

Table 12. Number of covariates (% of covariates with the standardized difference of means >0.1) for selected time intervals.

	Index date (day 0)		Long-term (450 to 181 days prior to the index date)	
	CUIMC	Optum EHR	CUIMC	Optum EHR
COVID-19 vaccinated patients compared to unvaccinated patients indexed on a date	9,073 (0.3%)	15,097 (<0.1%)	26,859 (0.5%)	51,075 (0.1%)
COVID-19 vaccinated patients compared to unvaccinated patients indexed on a visit	18,741 (2.2%)	21,739 (0.5%)	37,073 (0.1%)	50,358 (0.2%)
Influenza vaccinated patients compared to unvaccinated patients indexed on a date	12,684 (3.7%)	26,809 (0.1%)	25,782 (3.4%)	55,665 (0.4%)
Influenza vaccinated patients compared to unvaccinated patients indexed on a visit	22,816 (1.4%)	32,931 (0.1%)	34,361 (1.5%)	56,387 (0.2%)

Comparison of vaccinated patients and unvaccinated patients indexed on a date or a visit

Influenza vaccinated population

On the index date (day of vaccination = day 0), the influenza vaccinated population had markedly higher proportion of most covariates than an arbitrary date in the comparison group (pinning most covariates against the X-axis in Figure 23, A and B, yellow). The largest difference in covariate

proportions between unvaccinated and vaccinated populations on day 0 was observed for inpatient and outpatient measurements such as blood count, metabolic panels, blood pressure and basal metabolic index, including both presence of measurements and proportion of patients with of abnormal results, meaning patients were far more likely to have measurements on the date of vaccination than on an arbitrary date. Moreover, influenza vaccinated population had higher covariate proportions even a year prior to the vaccination.

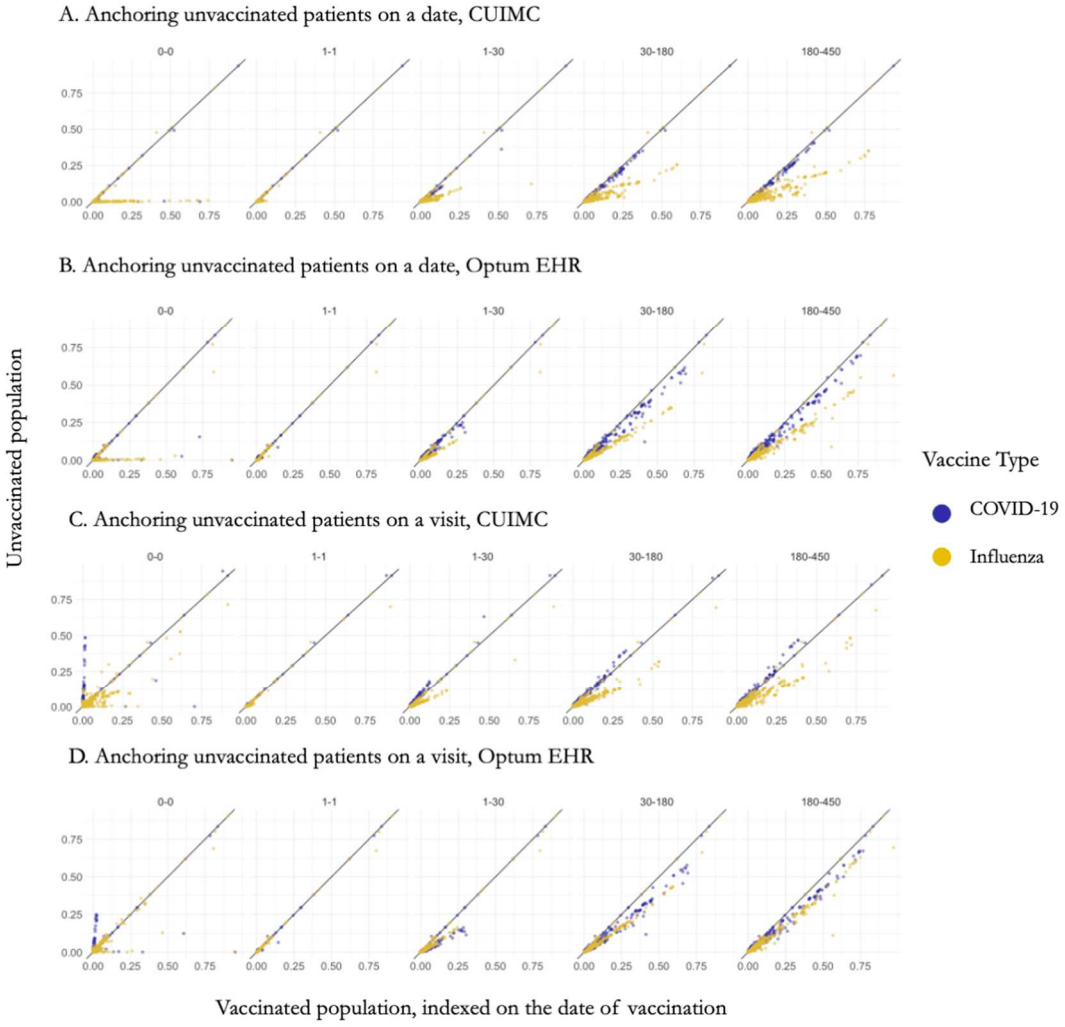


Figure 23. Baseline covariate proportion in vaccinated and unvaccinated populations on day 0, day -1, days -1 to -30, -31 to -180 and -181 to -450 in CUIMC and Optum EHR. Each dot represents a covariate; blue – covariate proportion in COVID-19 vaccinated population versus unvaccinated population and yellow – in influenza vaccinated population versus unvaccinated population.

In contrast, comparison with unvaccinated population indexed on a visit (Figure 23, C and D) showed a smaller difference between covariate proportions in CUIMC and almost no difference in Optum EHR, potentially indicating that a visit is a better counterfactual for a vaccination date than an arbitrary date.

Covariate proportions in vaccinated patients were closer to the proportions in the unvaccinated population indexed on a visit even with a longer lookback period.

COVID-19 vaccinated population

As opposed to the influenza vaccinated population, the difference in covariate proportion between COVID-19 vaccinated population and unvaccinated population indexed on an arbitrary date was moderate. We observed that COVID-19 vaccination was associated with a visit in 2.7% of patients (compared to 1.2% on an arbitrary date). In contrast, 55% of influenza vaccinated population had a visit on the date of vaccination (compared to 0.5% of unvaccinated population on an arbitrary date). Vaccinated population tend to have higher proportion of covariates prior to the index date (looking back a year prior).

When compared to the unvaccinated population indexed on a visit, COVID-19 vaccinated population had markedly lower proportion of most covariates. Those vaccinated with COVID-19 vaccine had much lower rates of diagnoses of both chronic and acute diseases on the date of vaccination

compared to a visit in unvaccinated population. The list of conditions included common chronic conditions such as hypertension, depressive disorder, asthma, and diabetes mellitus along with acute conditions like dyspnea, chest pain and fever. Such a difference points out that an arbitrary date may be a better counterfactual for a vaccination date in COVID-19 vaccinated patients.

Comparison of vaccinated patients indexed on the date of vaccination and the same patients indexed on a prior date or visit

Influenza vaccinated population

Here, we compared vaccinated patients indexed on the vaccination date to the same patients indexed on a date or visit within a year prior as we would do in a self-controlled study. We observed that the date of influenza vaccination tended to have a higher proportion of covariates compared to an arbitrary date within a year prior (Figure 24, first column) and even higher compared to an arbitrary visit within a year prior. Patients indexed on the date of vaccination were more likely to have antecedent healthcare encounters, conditions and laboratory tests within a year prior to the vaccination date than within a year prior to their previous visits (Figure 24, C and D). For comparison with an arbitrary date, we observed a mixed effect: in Optum EHR, vaccinated patients had more events preceding their vaccination while in CUIMC they had fewer events. Nevertheless, in both data sources the difference between covariate proportions was larger in magnitude when compared to an arbitrary date than when compared to an arbitrary visit.

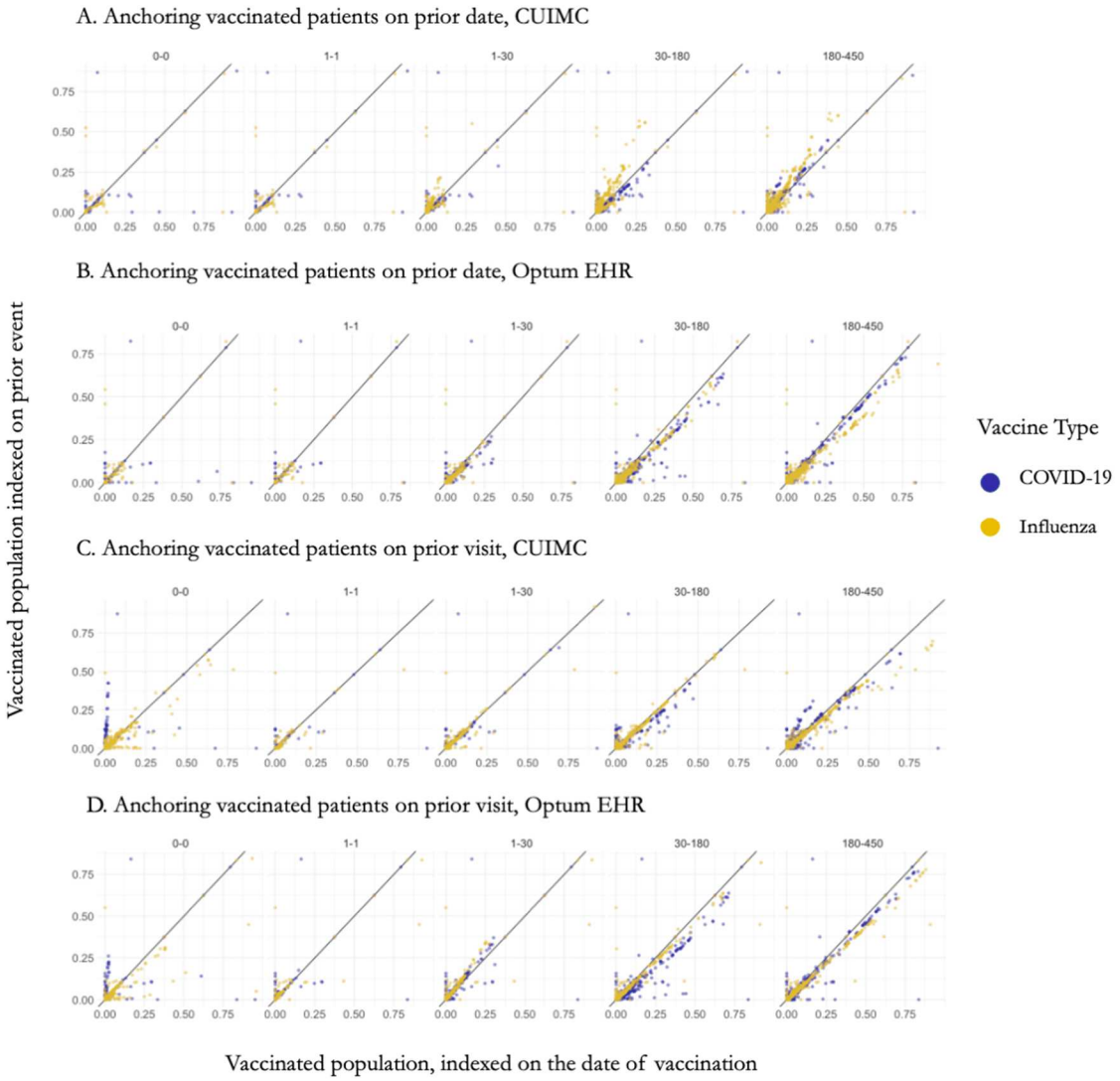


Figure 24. Baseline covariate proportion in vaccinated population indexed on the date of vaccination compared to the same population indexed on a prior visit or date on day 0, day -1, days -1 to -30, -31 to -180 and -181 to -450 in CUIMC and Optum EHR. Each dot represents a covariate; blue – covariate proportion in COVID-19 population and yellow – in influenza vaccinated population.

COVID-19 vaccinated population

The COVID-19 vaccinated population showed a markedly lower proportion of covariates on the day of vaccination compared to a visit or an arbitrary date within a year prior to vaccination. The difference was attenuated with a longer lookback period: COVID-19 vaccinated patients had fewer healthcare events within a year prior to their vaccination compared to their previous history. We observed mixed effect when comparing to a date in the past: some covariates such as exposure to COVID-19, COVID-19 laboratory tests, vital signs or acetaminophen were present in a higher proportion immediately before vaccination. Others like glomerular filtration rate, thyrotropin measurement, urinalysis or glomerulonephritis were observed in a lower proportion immediately before the vaccination.

3.1.2.4 Discussion

As we saw, patient baseline covariates in the unexposed group or time are extremely sensitive to the choice of the index date. We found that COVID-19 vaccination and influenza vaccination differ drastically from each other, with the proportion of most covariates much higher on the date of vaccination in the influenza group than the COVID-19 group. Therefore, study designs previously used to assess influenza vaccination must be reassessed for COVID-19 to account for a potentially healthier population and lack of medical activity on the day of vaccination. For other medical products, anchoring has to be carefully assessed and the index date should be chosen based on the knowledge about administration patterns and empirical covariate balance examination as described in this study.

Persons vaccinated for influenza appear to have more comorbidities and past procedures and measurements than the average population, even after adjusting for age and sex, and persons vaccinated for COVID-19 appear to have a lower proportion of most medical covariates than the average population after adjusting for age and sex. This may be explained if influenza vaccination is targeted to

sicker populations on average and if COVID-19 is targeted to the general public, which is healthier on average than those in our electronic health records [8,10].

The drastic effects on day 0—i.e., the day of vaccination and its comparison—are likely related to the context in which the vaccination is given. If the comparison is an arbitrary date in the person's record, then influenza vaccination has markedly higher covariate proportions, reflecting the association of the vaccination with a healthcare encounter. Moreover, such a trend (not observed for COVID-19 vaccine) was present even when comparing the date of influenza vaccination to the prior patient visits.

The abovementioned trends for COVID-19 vaccine were consistently observed in both data sources and the differences between the data sources were mainly related to the coding practices. For example, in CUIMC data COVID-19 vaccination was not associated with a visit but rather with a patient encounter. On contrary, COVID-19 vaccination in Optum EHR was associated with the providers entering 'Requires vaccination' and 'Vaccine Administration' in the system along with the codes for the vaccines. For influenza vaccination, the observed patterns were also consistent when comparing the vaccinated population to the unvaccinated population. When looking at the vaccinated patients immediately before the vaccination compared to an arbitrary date in the past the mixed effect observed can be attributable to continuous surveillance of such patients in CUIMC, which results in having higher healthcare utilization over an extended time period in the past.

The first implication of these results is that, when comparing vaccinated to unvaccinated patients or time, the anchoring event for unvaccinated comparator must be selected carefully. Previous research acknowledged that comparing unexposed and exposed patients in the context of vaccine safety and effectiveness surveillance may lead to between-person confounding due to non-comparable groups [16]. For example, as noted before for influenza, vaccinated and unvaccinated patients differ in co-morbidity

prevalence [17]. Nevertheless, even in the same population, the choice of the index date or event influences both baseline covariates and incidence rates of conditions following the index date. For COVID-19 vaccination, it appears that the comparison should not be purposely anchored on a health care visit unless it is a relevant vaccination subgroup (e.g., those vaccinated at hospital discharge).

Adjusting for confounding will be extremely important, as it appears unlikely that a comparison can be chosen perfectly, although the comparisons between the same participants looking a year prior led to the best equivalency for both influenza and COVID-19. Moreover, the difference in patient characteristics require robust selection of covariates for a propensity score model or outcome model as opposed to the current exposed vs unexposed COVID-19 vaccine cohort studies, which only use a limited subset of covariates in their propensity score model [18].

Alternatively, this may argue for a self-controlled study design [19], which mitigates the difference in patient characteristics. However, this design is also sensitive to anchoring (which is what happens on day 0 and around it) and carries other challenges such as accounting for differences in COVID-19 risk over time. For example, we observed that the time before vaccination is compared to the time before a visit in the past, the former time interval is characterized by higher prevalence of COVID-19 diagnosis and COVID-19 laboratory tests in both data sources as the previous visits mainly had occurred in 2020 before COVID-19 pandemic started.

This study has implications beyond using covariates for confounding adjustment. The day 0 results have direct implications for analyses of acute side effects like anaphylaxis that include day 0 because the side effect often occurs immediately. Any study of such short-term effects must directly account for anchoring to the context in which the vaccination is given. Furthermore, studies that compare effectiveness or safety among vaccines must account for differences in populations and in

vaccination context. For example, single-dose vaccines may be given preferentially to sicker patients who are unable to return for a second dose, such as those being discharged from the hospital.

3.1.2.5 Limitations

This experiment was performed on two EHR data sources. The findings may not be generalizable to other types of data sources or international data sources, so the empirical evaluation of baseline covariate balance may have to be performed on those data sources separately.

While CUIMC EHR is linked to the City and State Registries which ensures complete vaccination capture, Optum EHR may lack accurate vaccination status for some of the adults, which indicates that the covariate balance showed is only applicable to the captured adults

4.2. Assessing bias in defining an unexposed comparator for effectiveness research[6]

In previous sections we discovered that defining an unexposed comparator in safety research must account for bias arising from anchoring patients on a date or event as well as for heterogeneity stemming from selecting a specific study population, temporal window and season.

This section dives deeper in the bias in vaccine effectiveness studies. We conduct a robust analysis of short-term and long-term COVID-19 vaccine effectiveness accompanied by several secondary analyses and chart review to discover and attempt to mitigate selection and health-seeking behavior biases as well as confounding by severity and indication.

4.2.1 Background

Randomized clinical phase-3 trials have demonstrated high efficacy for the four most commonly used COVID-19 vaccines against symptomatic COVID-19 infection, ranging from 66.9% and 70.4% for Ad26.COV2.S (Johnson & Johnson–Janssen) and ChAdOx1 (Astrazeneca) to 94.1% and 94.6% for BNT162b2 (Pfizer–BioNTech) and mRNA-1273 (Moderna) vaccines [399–402]. Their rapid approval and widespread use require robust post-marketing studies that leverage large sample size, heterogeneous populations, and longer follow-up available in observational data.

Recent observational studies showed effectiveness similar to the randomized clinical trials (RCTs) across the globe, both test-negative and cohort designs [403–410], followed by studies across different patient populations, variants and regimens [411–415].

---

[6] This section is published in BMJ Open. The full citation for this publication is:

Ostropolets, A., Hripcsak, G. COVID-19 vaccination effectiveness rates by week and sources of bias: a retrospective cohort study. *British Medical Journal Open* (2022)

Nevertheless, multiple questions on study validity were raised, including vaccine status misclassification [416], matching vaccinated and unvaccinated populations [404] and addressing disease risk factor confounding and ascertainment bias [417–419].

One of such questions is COVID-19 vaccine effectiveness during the first two weeks following the first dose. Studies have shown contradicting results for Pfizer–BioNTech vaccine with estimates ranging from moderate effectiveness of 52% [401] to very high effectiveness of 92.6% [420]. Similarly, a recent study showed an unexplained high effectiveness of Janssen vaccine during week one [421]. Other studies simply excluded the first week(s) from the time-at-risk [407,411,422–424].

Week one lack of effectiveness was suggested as a metric for lack of confounding in the long-term vaccine effectiveness studies, but the reasons for high effectiveness and its impact on the validity of the conclusions regarding the overall effectiveness remain unclear [407].

Here, we aim to assess underlying bias associated with the use of observational data for short-term vaccine effectiveness and its impact on long-term vaccine effectiveness estimates. We employ large-scale propensity score matching and many negative controls to reduce bias, and leverage a range of secondary analyses as well as manual review of the COVID-19 infection cases in week one to examine health-seeking behavior of vaccinated and unvaccinated patients.

4.2.2 Methods

Study design

For our main analysis, we studied two mRNA vaccines (Pfizer-BioNTech or Moderna). The exposed group included patients indexed on the first dose of one of the corresponding vaccines with no prior COVID-19 infection and no previous exposure to other COVID-19 vaccines (Figure 25). For the unexposed group, we selected unvaccinated patients and set their index date to a date that matched the

index date of one of the exposed group participants. Both the exposed and unexposed groups had at least 365 days of prior observation and primarily resided in New York City according to their zip code. Patients who did not reside in New York were excluded from the study to ensure reliable vaccination data capture.

Outcomes of interest included a) COVID-19 infection defined as a positive COVID-19 test (reverse-transcriptase–polymerase-chain-reaction assay) or a diagnostic code of COVID-19 and b) COVID-19 hospitalization defined as an inpatient visit associated with a COVID-19 positive test or diagnosis within 30 days prior or during the visit. Upon further examination of the results, we added two other outcomes: a) COVID-19 positive test only and b) COVID-19 hospitalization associated with a positive COVID-19 test [425].

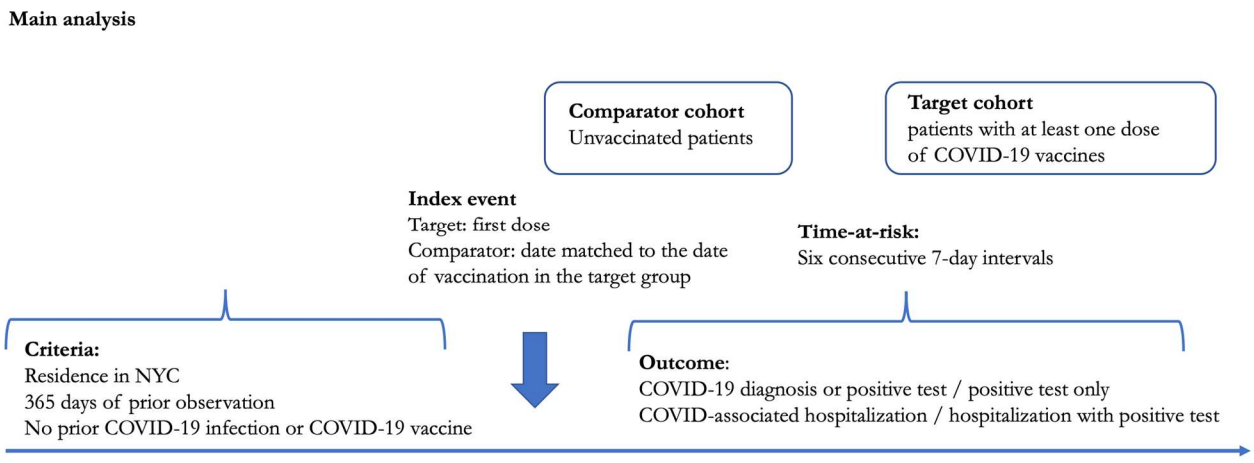


Figure 25. Overview of the design of the retrospective cohort COVID-19 vaccine effectiveness study.

We calculated vaccine effectiveness during six consecutive 7-day intervals after the first dose. Within each interval, patients were followed-up until an outcome, end of the period or death, whichever came earlier. Additionally, given the results for vaccine effectiveness during week one following the first dose, we conducted chart review for patients with a COVID-19 positive test recorded in the abovementioned period. We reviewed all cases for the vaccinated population as well a random sample of the cases in the unvaccinated population and extracted main complaint, COVID-19 history, including symptoms (fever, shortness of breath, sore throat, cough etc.), severity, time from the first symptom to encounter and COVID-19 exposure.

Secondary analyses

We also conducted a set of secondary analyses to examine potential biases. First, given that the published studies focused on patients without prior COVID-19 infection, we studied all eligible patients regardless of their previous COVID-19 status.

As the strategy for unvaccinated group index date selection (anchoring) has been reported to influence incidence of outcomes and baseline characteristics [10,11], we additionally tested unexposed patients indexed on a healthcare encounter matching the index date of one of the exposed group participants within 3 days corridor, with at least 365 days of prior observation located at New York.

Finally, we assessed vaccine effectiveness in patients with at least one dose of a COVID-19 vaccine and in fully vaccinated patients over all available follow-up to compare the estimates to the results of the RCTs. The latter was defined as 14 days after the second dose of Pfizer-BioNTech or

Moderna vaccines or first dose of Janssen vaccine. For each comparison we estimated hazard ratios (HRs) and constructed Kaplan-Meier plots as described below.

Statistical methods

For each analysis, we fitted a lasso regression model to calculate propensity score and match patients in each exposed and unexposed group with 1:1 ratio. For large-scale propensity score model we used all demographic information, index year and month, as well as the number of visits, condition and drug groups, procedures, device exposures, laboratory and instrumental tests and other observations over long (prior year) and short-term period (prior month) [207,212].

For each outcome, we fitted a Cox proportional hazards models to estimate HRs and constructed Kaplan-Meier plots. Empirical calibration based on the negative control outcomes was used to identify and minimize any potential residual confounding by calibrating HRs and 95% confidence intervals (CIs) [139,204]. Vaccine effectiveness was calculated as 100% × (1−hazard ratio).

Diagnostics

We used multiple sources of diagnostics to estimate potential bias following best practices for evidence generation [74]. First, we examined covariate and propensity score balance prior to proceeding with outcome modelling and effect estimation to ensure that we have enough sample size and to control for potential observed confounding [74]. We plotted propensity scores to investigate the overlap in patient populations at the baseline and examined the balance of all baseline characteristics to determine if the exposed and unexposed cohorts were imbalanced at the baseline and after propensity score matching. Exposed and unexposed cohorts were said to be balanced if the standardized difference of means of all covariates after propensity score matching was less than 0.1 [267].

For negative control calibration, we used 93 negative controls (Appendix 4.18) with no known causal relationship with the COVID-19 vaccines . Negative controls were selected based on a review of existing literature, product labels and spontaneous reports and were reviewed by clinicians [426]. We assessed residual bias from the negative control estimates.

We used CUIMC EHR (Table 1), which has an ongoing automated connection to New York City and State public health department vaccine registries and includes all within-state vaccinations for our population.

4.2.3 Results

Patient characteristics

In total, we identified 179,666 patients with at least one dose of COVID-19 vaccine in January-May 2021: 121,771 patients for Pfizer-BioNTech, 52,728 for Moderna and 5,167 for Janssen (Table 13).

The sample included patients from all age groups, with or without co-morbidities captured in inpatient and outpatient settings.

We observed that unexposed patients (Table 13) were on average younger and had fewer co-morbidities and less exposure to various drugs prior to matching. We were able to achieve balance on all covariates (up to 54,987 covariates, standardized difference of means less than 0.1) with propensity score matching. Appendix 4.19 presents the covariate balance and propensity score balance plots showing that anchoring unvaccinated patients on a date allowed us to achieve better balance compared to anchoring patients on a visit.

Table 13. Patient baseline characteristics for patients with at least one dose of a COVID-19 vaccine and the unexposed patients, before and after propensity score matching.

	Before matching			After matching		
Characteristic	Vaccinated	Unvaccinated	Std. diff	Vaccinated	Unvaccinated	Std. diff
Pfizer-BioNTech COVID-19 vaccine						
Patients, n	121,771	164,997		101,109	101,111	
Follow-up, days. Median (IQR)	107 (80 – 137)	104 (71-137)		107 (78-149)	107 (79-140)	
COVID-19 diagnosis or positive COVID-19 test, n				822	1355	
Positive COVID-19 test, n				231	786	
Age group, %						
10-19	4.2	10.8	-0.25	4.8	4.3	0.02
20-49	37.2	42.6	-0.11	40.3	40.1	0

50-64	23.9	20.3	0.09	23.6	23.7	0
65-74	18.8	12.6	0.17	15.8	16.6	-0.02
75-84	11.3	8.9	0.08	10.6	10.7	0
>84	4.1	3.8	0.02	4.2	4.1	0.01
Gender, %						
Female	63.7	57.8	0.12	61.4	62	-0.01
Race, %						
race = Asian	3.8	2.6	0.07	3.5	3.4	0.01
race = Black or African American	12.4	14.2	-0.05	12.6	12.2	0.01
race = White	40.5	35.1	0.11	39.3	39.5	0
Medical history, %						
Chronic liver disease	0.6	0.6	0	0.5	0.5	0
Chronic obstructive lung disease	1.3	1	0.02	1	1	0.01
Dementia	1.2	1.1	0	1.1	1	0.01
Depressive disorder	5.3	4	0.06	4	3.7	0.02

Diabetes mellitus	7.1	5.2	0.08	5.7	5.4	0.01
Human immunodeficiency virus infection	1.4	1.1	0.03	1.1	1	0
Hyperlipidemia	12.9	8.1	0.16	10.2	9.5	0.02
Hypertensive disorder*	16	11.3	0.14	13.1	12.2	0.03
Obesity	5.1	4.9	0.01	4.4	4.1	0.02
Osteoarthritis	7.3	4.7	0.11	5.8	5.3	0.02
Renal impairment**	3.7	3	0.04	2.9	2.7	0.01
Cerebrovascular disease	1.7	1.4	0.02	1.5	1.4	0.01
Heart disease***	8.6	7.1	0.06	7.5	7.1	0.02
Malignant neoplastic disease	5.3	4.5	0.04	4.7	4.3	0.02
Charlson comorbidity index, mean (SD)	1.75 (3.18)	1.69 (3.09)	-0.01	1.70 (3.11)	1.63 (3.03)	-0.01
Influenza vaccination within a year prior	10.9	7.9	0.10	7.5	6.9	0.02

Moderna COVID-19 vaccine						
Patients, n	52,728	148,795		50,517	50,517	
Follow-up, days. Median (IQR)	127 (102 – 153)	123 (99-153)		126 (101- 153)	126 (102-153)	
COVID-19 diagnosis or positive COVID-19 test, n				382	786	
Positive COVID-19 test, n				94	447	
Age group, %						
10-19	0.5	1.7	-0.12	0.5	0.4	0.01
20-49	35.7	45.7	-0.20	36.9	37.4	-0.01
50-64	21.2	23.3	-0.05	21.7	21.4	0.01
65-74	21.3	14.4	0.18	20.6	20.5	0.00
75-84	15.4	10	0.16	14.6	14.6	0.00
>84	5.8	4.8	0.04	5.6	5.6	0.00
Gender, %						

Female	64.4	58.7	0.12	64.2	64.7	-0.01
Race, %						
race = Asian	4.2	2.8	0.07	4.2	4.4	-0.01
race = Black or African American	8.7	14.2	-0.17	9	8.4	0.02
race = White	48.3	34.4	0.29	46.9	47.9	-0.02
Medical history, %						
Chronic liver disease	0.5	0.6	-0.02	0.5	0.5	0
Chronic obstructive lung disease	1.4	1.1	0.02	1.2	1.2	0
Dementia	1	1.2	-0.02	1	0.9	0.01
Depressive disorder	4.7	3.9	0.04	4.2	4	0.01
Diabetes mellitus	6.6	5.6	0.04	6.2	5.8	0.02
Human immunodeficiency virus infection	0.9	1.2	-0.03	0.8	0.8	0
Hyperlipidemia	14.9	8.9	0.19	13	12.6	0.01
Hypertensive disorder	16	12.4	0.1	14.7	13.9	0.02

Obesity	4	4.4	-0.02	3.8	3.6	0.01
Osteoarthritis	7.7	5.3	0.1	6.8	6.5	0.01
Renal impairment	3.5	3.3	0.01	3.3	3	0.01
Cerebrovascular disease	2.2	1.6	0.05	2	1.8	0.02
Heart disease	10.1	7.6	0.09	9.2	8.7	0.02
Malignant neoplastic disease	6.5	5	0.07	5.9	5.5	0.02
Charlson comorbidity index, mean (SD)	1.62 (2.81)	1.62 (3.00)	0.00	1.59 (2.80)	1.59 (2.99)	0.00
Influenza vaccination within a year prior	8.4	6.3	0.08	7.2	6.8	0.02
Janssen COVID-19 vaccine						
Patients, n	5,167	52,643		5,031	5,031	
Follow-up, days. Median (IQR)	79 (72-95)	79 (72-95)		79 (72-95)	79 (72-95)	
COVID-19 diagnosis or positive COVID-19 test, n				31	37	

Positive COVID-19 test, n				8	16	
Age group, %						
10-19	0.8	0.8	0.00	0.8	0.8	0.00
20-49	43.9	43	0.02	44.2	43.9	0.01
50-64	31.7	31.7	0.00	31.8	31.3	0.01
65-74	11.6	12.2	-0.02	11.5	12	-0.02
75-84	7.6	7.9	-0.01	7.2	7.9	-0.03
>84	4.3	4.3	0.00	4.2	4	0.01
Gender, %						
Female	63.4	63.2	0.01	63.5	61.1	0.05
Race, %						
race = Asian	3.6	1.7	0.12	3.7	3.6	0.01
race = Black or African American	15.9	15.5	0.01	15.7	15.5	0
race = White	37.4	35.7	0.03	37.4	37.5	0
Medical history, %						

Chronic liver disease	1.1	0.7	0.05	1	1.2	-0.02
Chronic obstructive lung disease	2.4	1.3	0.09	2	2.2	-0.01
Dementia	2.6	1.1	0.11	2.2	2.2	0
Depressive disorder	8	4.8	0.13	7.1	8	-0.03
Diabetes mellitus	10.3	6.2	0.15	9.5	10.2	-0.02
Human immunodeficiency virus infection	1.7	1.4	0.02	1.6	1.8	-0.01
Hyperlipidemia	14.3	10.2	0.13	13.4	14.3	-0.03
Hypertensive disorder	21.4	13.8	0.2	20.1	21.7	-0.04
Obesity	7.3	5.9	0.06	6.8	7.8	-0.04
Osteoarthritis	8.4	6.2	0.08	7.8	8.8	-0.04
Renal impairment	6.6	3.3	0.15	5.3	5.9	-0.02
Cerebrovascular disease	2.7	1.7	0.07	2.3	2.4	-0.01
Heart disease	11.8	8	0.13	10.3	11.7	-0.04
Malignant neoplastic disease	5	4.9	0	4.8	5.2	-0.02

Charlson comorbidity index, mean (SD)	1.84 (3.34)	1.55 (2.96)	-0.07	1.56 (3.04)	1.43 (2.79)	-0.03
Influenza vaccination within a year prior	12.5	8.0	0.15	10.1	11.4	-0.04

* Hypertensive disorder includes primary and secondary hypertension

** Renal impairment includes acute and chronic renal failure (prerenal and renal);

*** Heart disease includes cardiac arrythmias, heart valve disorders, coronary arteriosclerosis, heart failure, etc.

Patients vaccinated with Pfizer-BioNTech had a similar distribution of baseline characteristics compared to the patients vaccinated with Moderna but differed from the patients vaccinated with Janssen. On average, the latter group was older, had more patients with race recorded as Black, and had more co-morbidities such as diabetes mellitus or hypertensive disorder.

When investigating the vaccination pathways, we discovered that 112,963 patients (93% of patients with at least one dose of Pfizer-BioNTech) had 2 doses of Pfizer-BioNTech and 42,384 (80%) patients had 2 doses of Moderna. We found 344 and 291 patients with 3 doses of the corresponding vaccines and 440 patients having mixed Pfizer-BioNTech, Moderna and Janssen vaccines in different combinations.

Within our database, Moderna was administered early on with a peak in January 2021 (Figure 26), while Pfizer-BioNTech and Janssen vaccinations peaked in April. It was reflected in the follow-up time with Moderna patients having on average longer follow-up with some individuals having up to 5.8 months of post-observation.

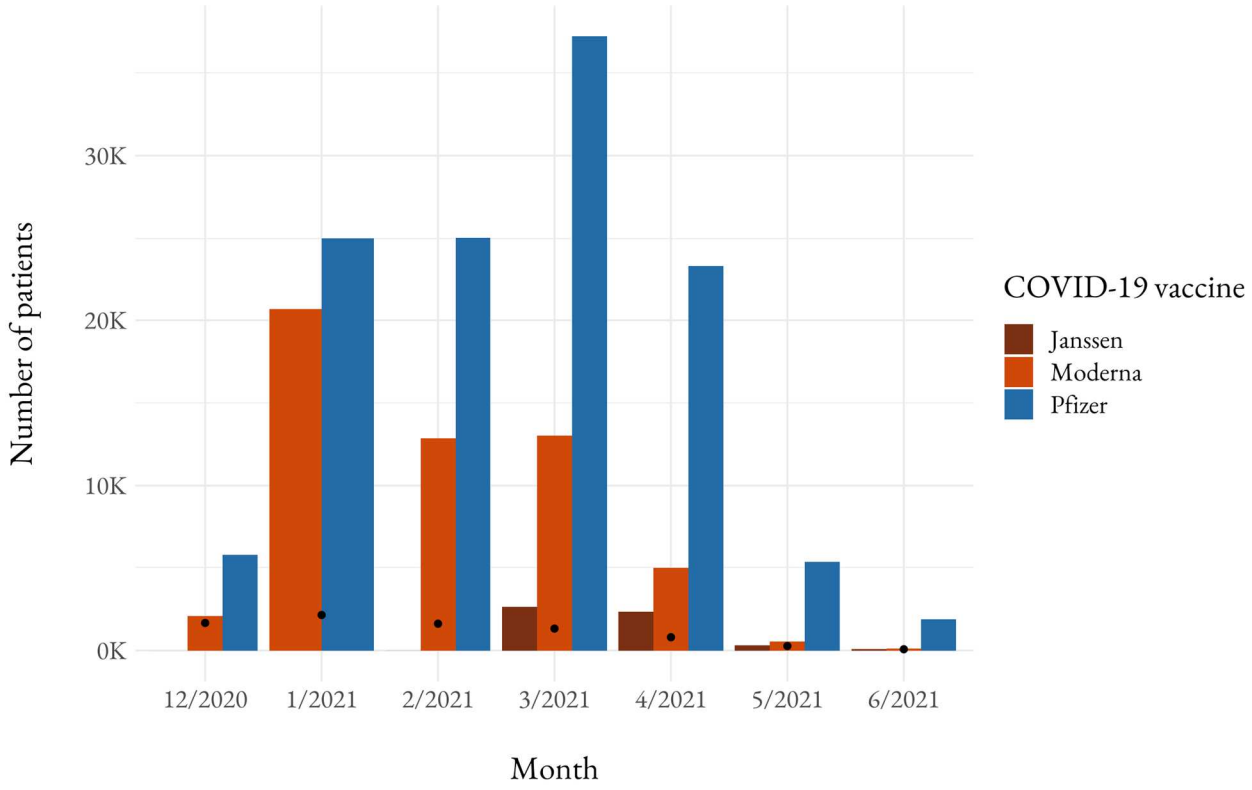


Figure 26. Distribution of vaccination month for COVID-19 vaccines. Black dots represent the number of incident COVID-19 cases (defined as a positive test) in each month.

Main week-by-week effectiveness analysis

Figure 27 shows vaccine effectiveness over six 7-day intervals for patients vaccinated with at least one dose of Pfizer-BioNTech or Moderna (160,114 patients) compared to unvaccinated patients (115,689). Due to the small sample size, we were not able to obtain stable week-by-week estimates for Janssen.

While week one was characterized by unexpectedly high effectiveness (58%, 95% CI 45-69% against COVID-19 infection and 72%, 95% CI 57-83% against COVID-19 associated hospitalization), we observed plausible increasing effectiveness beginning week 2 with the effectiveness on week 6

approximating 84% (95% CI 72-91%) for COVID-19 infection and 86% (95% CI 69-95) for COVID-19-associated hospitalization.

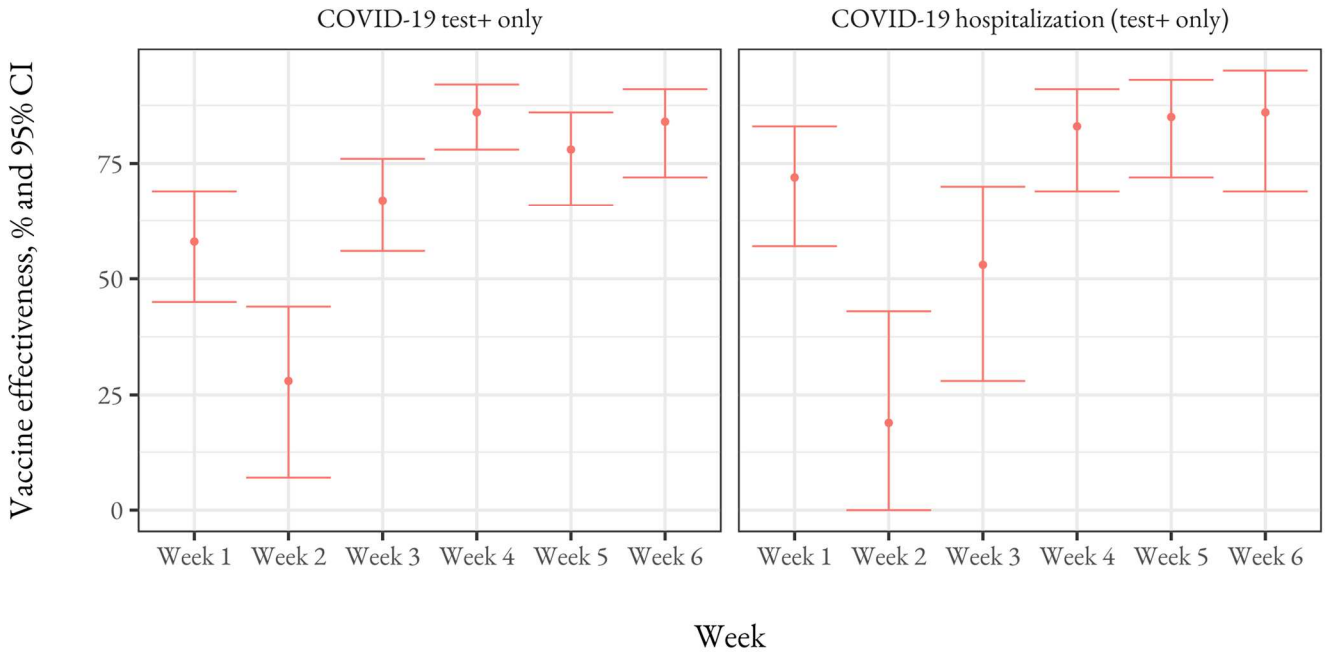
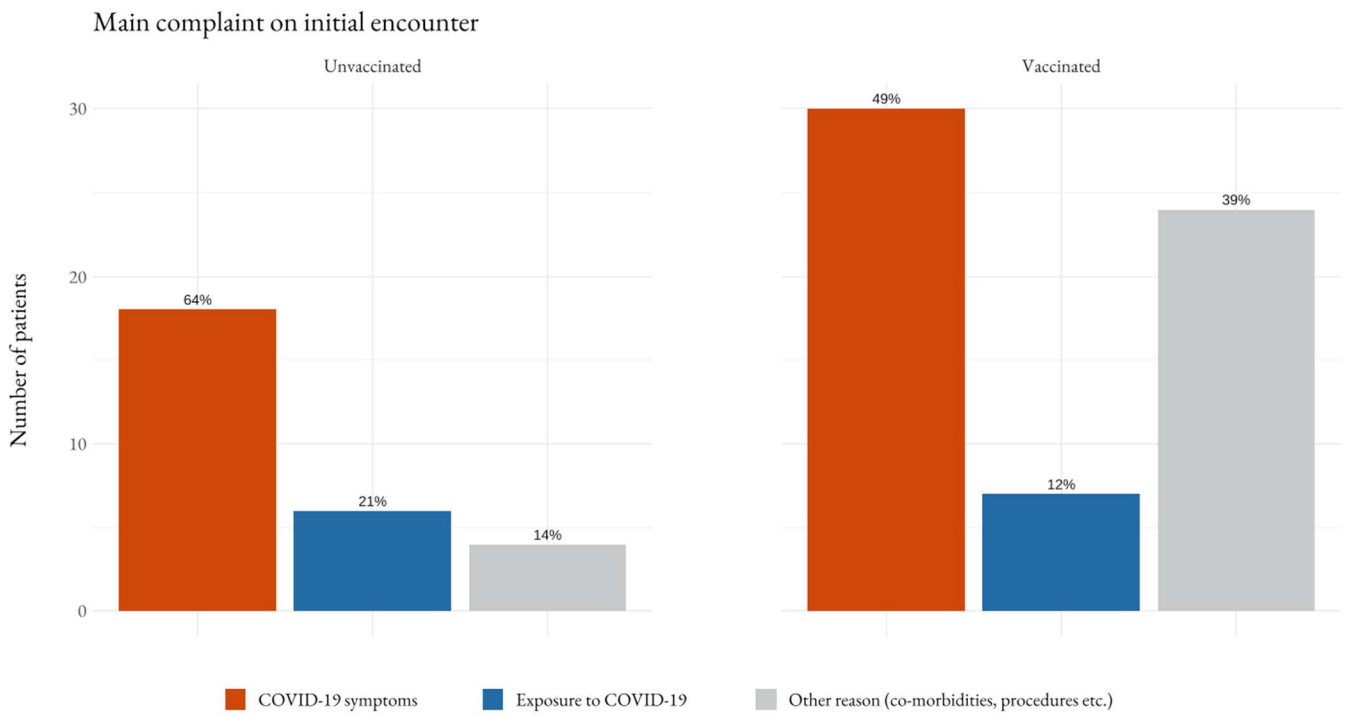


Figure 27. Effectiveness of Pfizer-BioNTech and Moderna vaccines over six 7-day intervals after 1st dose, % and 95% CI for COVID-19 infection and COVID-19 hospitalization.

We then looked at the week one COVID-19 infection cases to explain high effectiveness (Figure 28). A chart review of week one positive COVID-19 tests revealed a high proportion of unvaccinated patients seeking care related to COVID-19 symptoms or COVID-19 exposure (85% in total) compared to only 69% of vaccinated patients. Initial healthcare encounters in vaccinated population were oftentimes related to other medical reasons such as co-morbid conditions or surgeries (39% compared to 21% in unvaccinated population, Appendix 4.20). Moreover, an observed gap between symptom onset and an initial healthcare encounter was more pronounced in the vaccinated cohort as the patients attributed their symptoms to temporal vaccine side effects as opposed to COVID-19 infection.


Main complaint on initial encounter
Unvaccinated
Vaccinated
Number of patients
30
20
10
0
64%
21%
14%
49%
12%
39%
COVID-19 symptoms
Exposure to COVID-19
Other reason (co-morbidities, procedures etc.)

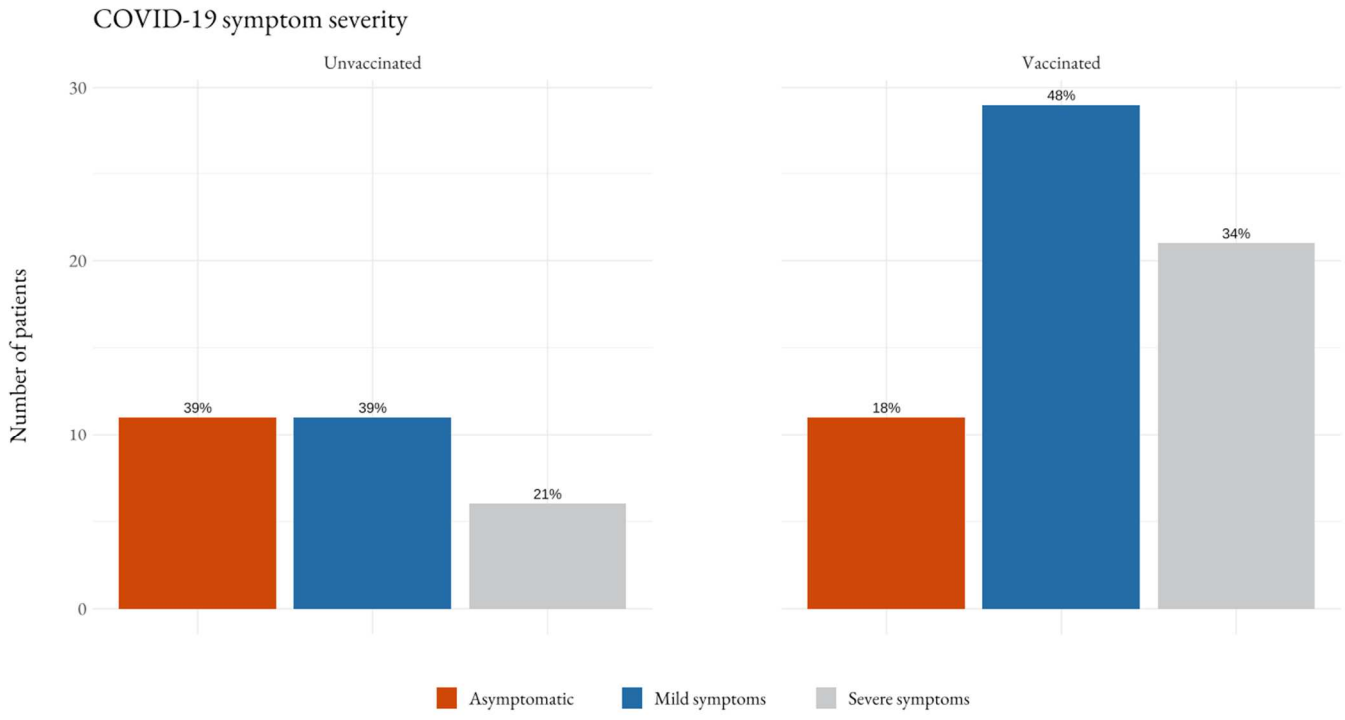

COVID-19 symptom severity
Unvaccinated
Vaccinated
Number of patients
30
20
10
0
39%
39%
21%
18%
48%
34%
Asymptomatic
Mild symptoms
Severe symptoms

Figure 28. Chart review of COVID-19 cases (defined as a positive COVID-19 test) during week one, vaccinated and unvaccinated patients, main complaint on initial encounter (top) and COVID-19 symptom severity (bottom).

When looking at the severity of COVID-19 symptoms at the initial encounter during week one after the index date, we observed that the unvaccinated cohort had a higher proportion of asymptomatic cases (39% compared to 11%) while the vaccinated population had more severe or mild cases (34% and 48% respectively).

Secondary analysis

As cohort analysis allows us to construct Kaplan-Meier curves to assess effectiveness over time, we also looked at the effectiveness during the year after the first dose (Appendix 4.21-4.23). We observed similar trends with all three vaccines being less effective during the first month after the first dose. After that, Pfizer-BioNTech and Moderna were highly effective against both COVID-19 infection and COVID-19 associated hospitalization, while Janssen vaccine exhibited a wide range of effectiveness (Appendix 4.24).

The results for fully vaccinated patients with time-at-risk starting at the full vaccination matched the results of the clinical trials for corresponding vaccines (detailed estimates are provided in Appendix 4.25 and 4.26).

Our initial design included a positive COVID-19 test or a diagnostic code as an outcome. Upon further case examination, we discovered that COVID-19 diagnostic codes in the CUIMC data were partially assigned to the patients with negative COVID-19 tests on or immediately following the date of diagnosis. In that case, ICD-10(CM) code U07.1 “Disease caused by Severe acute respiratory syndrome coronavirus 2” was entered in the system for billing purposes (COVID-19 molecular or antibody tests)

or for COVID-19 sequelae. We, therefore, focused on positive COVID-19 test only for our primary outcome, which led to higher effectiveness for all vaccines compared to using both positive test and diagnosis.

Finally, exclusion of patients with prior COVID-19 infection in our main analysis resulted in higher effectiveness. Inclusion of patients regardless of their prior COVID-19 status led to a small decrease in observed effectiveness (Appendix 4.27) for both COVID-19 infection and hospitalization in patients vaccinated with Moderna or Janssen.

4.2.4 Discussion

In this retrospective cohort study, we examined the effectiveness of COVID-19 mRNA vaccines over six 7-day intervals after the first dose. We scrutinized the effectiveness of the mRNA vaccines following the first dose and confirmed the findings of moderate vaccine effectiveness during the first two weeks. For week one following the first dose we discovered previously uncaptured differential biases in vaccinated and unvaccinated populations resulting in high vaccine effectiveness. Other researchers suggested that the difference between vaccinated and unvaccinated groups can be mitigated by adjusting for previous healthcare utilization such as number of visits before baseline, co-morbidities or prior vaccination behavior [404,411,422]. Nevertheless, the confounding we observed remains even upon controlling for a large number of covariates including those above.

Vaccination directly influenced the attitude of patients towards their symptoms, causing a delay in seeking care and a higher symptom severity threshold needed to seek care or get tested. On contrary, vaccinated patients in other studies had higher rates of testing compared to unvaccinated [417,427]. This indicates that patients' attitude toward risk of infection and testing may vary geographically and over time. Similarly, frequency of testing may depend on local policies and practices.

In unvaccinated patients, mild COVID-19 related symptoms were the reason to seek care; in vaccinated patients such cases were mainly captured upon seeking outpatient and inpatient care for other conditions.

For example, vaccinated patients could be hospitalized for elective surgery or delivery and be tested positive for COVID-19 on the day of admission or later on. Differential symptom severity was previously reported for other vaccines [428] and may affect any observational study that uses hospitalization as a surrogate for COVID-19 severity as it can be hard to accurately identify the main reason for hospitalization in structured data.

Previous research suggested that vaccinated patients do not have an increase in the number of cases immediately following vaccination as they are unlikely to get vaccinated if sick [123,407]. Our review of the cases in week one adds to 'healthy vaccinee' effect by showing that vaccinated patients are more likely to attribute their symptoms to common vaccine side effects and, therefore, are less likely to seek care.

Nevertheless, even when this differential bias is present, the estimates of the COVID-19 vaccine effectiveness in subsequent weeks still match the results of the RCTs. This indicates that high effectiveness during week one following vaccination does not necessarily undermine the estimates of subsequent vaccine effectiveness. On the other hand, we argue against using estimates of vaccine effectiveness within a short period after the vaccination as a negative control as the differences between the groups observed in this study are likely to be time-variant and may diminish over time [429].

Our secondary analyses discovered several challenges and potential biases that must be accounted for when conducting vaccine effectiveness studies on observational data. First, we observed that outcome definitions are prone to measurement error, which has not been studied thoroughly. Some

of the published studies used ICD-10 or ICD-10(CM) codes to identify COVID-19 outcomes [430–432]. We found that the specifics of data capture and billing processes were associated with some patients having assigned COVID-19 diagnosis codes for billing for tests rather than as an indicator of active disease. Another reason for assigning the code was COVID-19 sequela, where the actual date of COVID-19 infection could have been anywhere from 6 months to a couple of weeks in the past. Some researchers have previously reported high positive predictive value of ICD-10 diagnostic codes for COVID-19, which points out that index date misclassification should be scrutinized in each institution participating in the analysis to make valid inferences [433,434].

Second, inclusion or exclusion of patients with prior COVID-infection influenced estimated effectiveness. We observed that inclusion of patients with prior COVID-19 leads to lower effectiveness for all vaccines regardless of the outcome definition.

Third, an appropriate index event (anchor) for the unvaccinated cohort must be chosen to represent a counterfactual for vaccination [10,435]. In our study, we confirmed that an arbitrary date represents a better counterfactual than a medical visit for COVID-19 vaccination, which is reflected in propensity score balance and covariate balance. Nevertheless, other institutions may have different vaccination pathways such as vaccination on discharge, which can make a visit a better counterfactual for vaccination. More generally, completeness of vaccination data capture is a crucial feature that influences the robustness of the study. While CUIMC data ensures complete exposure capture by linking EHR to the City and State Registries, the researchers should exhibit caution with conducting studies on the data sources with unknown vaccination capture.

In general, our findings support the RCTs and previously published post-marketing studies for all three vaccines. Larger sample size for patients vaccinated with COVID-19 mRNA vaccines allowed us

to have more power, which resulted in overlapping yet narrower confidence intervals compared to the RCTs. On the other hand, our study had fewer patients with the Janssen vaccine, which resulted in wider yet overlapping intervals compared to the Janssen's vaccine RCT [399,400,405]. Nevertheless, an indirect comparison of these vaccines may not be accurate due to the differences in the populations we observed in our study. First, patients vaccinated with Janssen were substantially different from mRNA patients: on average, they were older, had a higher proportion of patients with race recorded as Black and had more comorbidities. Therefore, comparative effectiveness studies of Janssen and mRNA vaccines require robust techniques such as large-scale propensity matching to ensure valid comparison.

Second, while Moderna and Pfizer patients had similar baseline characteristics, the temporal distribution of vaccinations in CUIMC data differ. Moderna vaccine was administered early on in 2021 with the peak in January, while Pfizer vaccination peaked in April.

Given the fluctuating baseline COVID-19 infection prevalence in 2021, a comparison of mRNA vaccines requires matching patients on calendar month to account for this potential bias. These vaccines also had different administration pathways in our system. As opposed to Pfizer vaccine, which was administered at CUIMC/NYP sites to all patients over a prolonged period, Moderna vaccination was performed elsewhere and recorded for actively observed patients. Such patients were more likely to get tested or receive care outside of our healthcare system.

4.2.5 Limitations

Due to observational nature of the study, the data sources may not have complete capture of patient conditions as the patients could seek care outside of the hospital system. While our outcome phenotype algorithms may be subject to measurement error, we provided additional analyses with alternative outcome definitions. Exposure misclassification was mitigated by having free and available

COVID-19 testing and COVID-19 vaccination in Columbia University Irving Medical Center and New York-Presbyterian sites as well as by having data capture from New York City and State Immunization Registries. Along with availability of testing, COVID-19 baseline infection rate difference was mitigated by matching the exposed and unexposed groups on the index date and using the index month as a covariate in propensity score model. We attempted to address potential differences between exposed and unexposed groups by selecting a large number of covariates in our propensity score model such as number of visits, procedure and drug utilization, prior vaccine behavior, race and others.

The results of the study may not be generalizable to other countries or settings with different vaccine administration practices and policies. Finally, the study period did not allow us to stratify the results by COVID-19 variants, which limits generalizability of findings to other variants.

4.3 Chapter summary and lessons learned

This chapter uncovered several biases related to the choice of unexposed group and its index date both in safety and effectiveness research.

As we demonstrated, background rates commonly used in drug surveillance are sensitive to the choice of age group (variability in rates up to a factor of 1,000), index event definition or anchoring (up to a factor of 100) and data source (up to a factor of 10) and less sensitive to the choice of gender, race, season, or year. Anchoring also influences patient baseline characteristics with the effect being most prominent on the index date but persisting up to a year prior.

To ensure reliability of observed-to-expected analysis, we recommend using one data source for both cohorts' construction, ensuring that the exposed and unexposed have similar demographics distribution, time-at-risk and are captured within the same year and season. Additionally, the researchers must ensure that appropriate methods for confounding adjustment are used, arguing for large-scale propensity score matching as one of the robust methods that uses the maximum number of covariates possible to balance the cohorts. As self-controlled case series is also prone to anchoring, similar procedures can be used to balance the exposed and unexposed time [436]. We recommend empirically examining anchoring to select an appropriate index date or event based on the covariate distribution and background knowledge about drug administration settings.

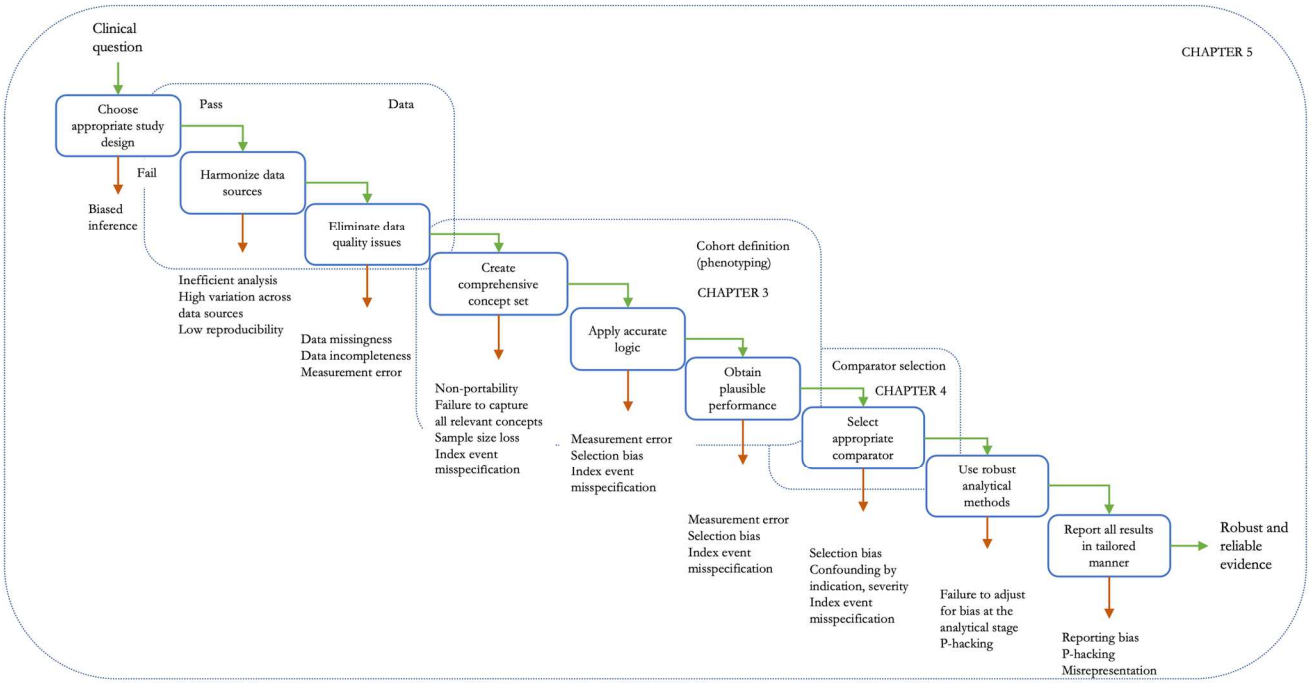
We observed that short-time (e.g., two weeks) estimates are especially prone to bias. Fist, the estimates of background rates of anaphylaxis, which were estimated within a short time-at risk, were sensitive to demographics, anchoring, season, year, and new case identification strategy. Similarly, short-term effectiveness estimates were prone to bias despite the robustness of long-term estimates. There were several biases that the adjustment on the analysis stage failed to correct for: selection bias,

bias by severity and indication and healthcare seeking bias. Measurement error discussed in detail in the previous chapters may have larger influence on short-term estimates if the number of events is small as demonstrated when comparing the diagnosis-based COVID-19 phenotype and laboratory test-based one.

Given that these biases were uncovered only through in-depth chart review of cases, we suggest that any study should thoroughly examine patient characteristics and use extensive diagnostics, especially if the estimates deviate from the baseline assumptions or background knowledge. In future, the formal and standardized processes should be established.

Chapter 5. Generating responsive and reliable evidence at the point of care

RCTs cannot readily address all the clinical questions and a substantial number of them remains unanswered [13,43,49–52]. Upon creating standardized procedures, frameworks and tools for generating reliable evidence, the process can become less time-consuming, which may enable evidence generation in real or near-real time. This chapter builds up on the knowledge learned in the previous chapters to examine the ability of observational data to support evidence generation and delivery at the bedside. Throughout this section we hypothesize and show that robust evidence can be delivered on-demand at the point of care.



We first interview 30 clinicians at NewYork-Presbyterian hospital, use thematic analysis and develop a taxonomy of information needs related to a gap in knowledge unmet by the current evidence.

We demonstrate that despite the abundancy of knowledge, clinicians have multiple questions that are not covered by the current evidence and oftentimes have shared areas of unmet needs such as optimal treatment of patients with multiple comorbidities or rare disorders, elderly and children or effectiveness and safety of new drugs. We discover that experienced and inpatient physicians would benefit from a clinical decision support system (CDSS) that generates evidence at the point of care.

We proceed by conducting a formal scoping review of clinical decision support tools that generate real-world evidence, which identifies 25 expert-based and data-driven tools that can generate new clinical evidence and further classify them according to their approach to evidence generation. We discover that the tools oftentimes are on the prototype stage and lack demonstration of their utility so that their impact on healthcare processes and patient outcomes remains unclear. Moreover, thorough and systematic assessment of bias in such CDSSs is lacking.

Based on these discoveries, we design and pilot a study of a data consult service that generates clinical evidence at the bedside with the clinicians affiliated with CUIMC. We create and implement a pipeline (question gathering, data exploration, iterative patient phenotyping, study execution, and evidence validity assessment) for generating new evidence in real time, which results in 24 answered questions collected from 22 clinicians.

We identify the key components required for successful early-stage implementation such as proactive involvement of the study team and participation in clinical rounds and shadowing. We classify and describe in-depth the main challenges we encountered such as missing and incomplete data, underreported conditions, nonspecific coding and accurate identification of drug regimens, and discuss user engagement and satisfaction.

5.1. Clinicians' information needs unmet by the current evidence [7]

As we hypothesize that the lessons we learned in Chapters 3 and 4 and the methods and pipelines we developed can enable responsive and reliable evidence generation, we explore the potential clinicians' questions that newly generated evidence can address. We conduct a series of interviews with clinicians at CUIMC and discover that, despite the availability of the evidence and growing number of observational studies, clinicians on average have at least four unmet evidence-related information needs, and such needs do not decrease with years of clinical experience accumulated.

The interviews result in a taxonomy of information needs including the areas commonly inadequately covered by the RCTs such as treatment of patients with multiple comorbidities or rare disorders, elderly and children, or questions related to effectiveness and safety of new drugs. We find that the most common topic is drug effectiveness, which aligns well with the earlier studies, indicating that it may be the most common use case for the service that delivers evidence at the bedside. We observe that specialty physicians seek more in-depth knowledge, which indicates that they, along with the experienced clinicians, are likely to benefit from such a service.

5.1.1 Background

Current evidence has been reported to be inadequate or missing for specific clinical cases [40,82]. For example, guidelines include a large proportion of recommendations based on expert opinion or case studies indicating a lack of reliable data to act upon [437] and cannot always address physicians' information needs.

---

[7] This section is published in JAMIA Open. The full citation for this publication is:
Ostropolets A, Chen R, Zhang L, Hripcsak G. Characterizing physicians' information needs related to a gap in knowledge unmet by current evidence. *JAMIA Open* (2020)

Apart from insufficient evidence, general information needs have been studied since the early stages of electronic health record systems adoption. Smith et al. in 1996 summarized the studies related to doctors' information needs, concluding that the prevailing part of the unmet information needs consists of treatment questions that are often complex and highly patient-specific [52]. This finding was supported by Ely et al., who found that most of the immediate questions generated during consultations remain unanswered, mainly due to the lack of time [82,438,439]. These studies did not distinguish between the information needs that could potentially be answered using existing evidence and those, for which no RTCs or clinical guidelines existed. While evidence exists, clinicians tend to use it to guide their decision-making [440,441]. Nevertheless, it is unclear which part of the clinical questions remained unanswered due to evidence lacking.

After more than twenty years, a growing pool of evidence requires an up-to-date assessment of the current evidence utilization and its ability to cover physicians' needs. More recent studies have mainly focused on specific cohorts of doctors: primary care physicians [442–445], family physicians [446,447] and residents [448–450], while the other specialties have not been studied thoroughly. Meanwhile, specialty physicians usually face complex clinical cases that may shape additional information needs [451] and remained unaddressed by the existing evidence. These studies addressed the immediate information requests mainly in outpatient settings where physicians have limited time and resources to answer their questions [82,438,442,443,452]. Surprisingly, there is little knowledge on evidence utilization and sufficiency among specialty physicians or senior physicians working in secondary and tertiary care services.

In this section, we address this knowledge gap, shedding light on current physicians' information needs that are specifically related to insufficient evidence and therefore cannot be answered with the

existing guidelines. These needs are not only of theoretical significance: with the growing pool of healthcare data, it has become possible to drive its secondary utilization to guide clinical decision-making by incorporating clinical decision tools into practice. Nevertheless, few of them have been adopted in routine practice. Among the other reasons, these tools might have been designed without a precise study of the current information needs and the proper target group identification. Unmet information needs specify user scenarios and use cases, which drives the methodology behind such CDSS, their design and implementation.

5.1.2 Methods

Unmet information needs

We used convenience sampling method to select thirty physicians (thirty participated, eight declined participation) affiliated with NewYork-Presbyterian and Columbia University Irving Medical Center. The one-hour in-depth face-to-face interviews were conducted over a 4-month period in the location of physicians' choice; one interview was conducted over the telephone. The open-ended semi-structured questionnaire included questions about (a) perceived information needs and (b) knowledge resources that physicians use to fulfill these needs (textbooks, electronic resources such as PubMed and commercial tools, clinical consultants, and pharmacists). To examine information needs, we asked interviewees to provide the number of questions for which they found no appropriate or insufficient medical evidence, the time expended searching for evidence, and examples of the questions that occurred. We provided examples of questions from our practice as well as other participants' scenarios. To facilitate recall, we provided scenarios related to different aspects of care (diagnosis of rare events or disorders, treatment strategies, quality of care, patient compliance).

We collected demographic data, number of years in practice, clinical rank, departmental affiliation, and clinical specialty. Physicians were additionally classified as practicing in an inpatient or outpatient setting based on the primary type of care they provide.

Thematic analysis

The interviews were analyzed by two independent investigators using deductive thematic analysis according to six phases defined by Braun and Clarke [453].

First, we entered the transcripts into the N-Vivo data management program with manual semantic code identification, which was moderated to ensure their validity and consistency across all transcripts. By creating codes applicable to all the data, we reduced the volume of text for analysis and identified new structures to incorporate disparate responses.

We applied cutting and sorting method [454] to identify the low-level codes and then generated a coding framework to connect related codes together to discover themes that are closely related to physicians' information needs. Merged codes formed themes that served as a basis for our themes of information needs related to a gap in knowledge unmet by current evidence. Each theme was refined to ensure proper association of its definition and name with the coded extracts and coherency across all interview transcripts.

The final themes were adjusted based on the discussions with both interviewees and investigators. Throughout the analysis, three co-researchers used pair debriefing to discuss disagreements and reach an agreement on the themes and subthemes achieving at least 80% inter-rater agreement between the investigators

Statistical analysis

We conducted additional statistical analyses to identify the correlation between clinical experience, clinical specialty, setting of clinical care and the characteristics of information needs. As we intended to study clinical information needs, we excluded two interviewees as they were not engaged in active clinical practice. To approximate perceived information needs, we used the number of questions or clinical scenarios that physicians recall during interviews and three main themes of clinical needs (diagnosis, treatment, and public health and quality of care) that we identified based on our hierarchical set of themes.

We then compared the distribution of the types of questions across specialties working settings and clinical experience using Pearson's Chi-squared test. To assess if there was a difference in the information needs among specialized and non-specialized physicians and among physicians working in different settings, Wilcoxon rank sum test with continuity correction was used; Kruskal-Wallis rank sum test was used to test if there is a difference in information needs based on clinical rank or departmental affiliation.

5.1.3 Results

We interviewed thirty physicians from a broad range of specialties: pediatrics (23.3%), general internal medicine (20%), nephrology (16.7%), cardiology (13.3%), neurology (6.7%), gastroenterology (6.7%), infectious diseases (6.7%), emergency medicine (3.3%), and intensive care (3.3%). On average, physicians have spent 13.4±8.3 years in clinical practice and most of them have an academic appointment (77%) : assistant (52%), associate (39%) or full professor (9%). Twenty-three percent of clinicians indicated outpatient practice as their main working setting and seventy-seven predominantly practiced in inpatient settings. Clinicians raised on average 4.3±2 questions per interview.

Thematic analysis

Twenty-seven physicians in our study said that they use evidence knowledge sources routinely. The other three listed socio-economic determinants, patient compliance and evidence irrelevance to the real-world practice as the obstacles to evidence use. Low patient compliance was said to influence medication prescribing and duration of therapy. For example, shorter courses of treatment or modified drug regimen were sometimes more convenient for patients but deviated from the current standards of care. Low income and hospital remoteness were also said to influence patterns of prescribing, favoring aggressive and short therapy. Some clinicians identified clinical guidelines and clinical trials as 'overcautious' and prescribed formally contraindicated drugs as they had not observed listed adverse events in their practice.

Ninety-one distinct clinical problems related to absent or insufficient medical evidence were raised during the interviews. The final taxonomy comprised fifteen end leaves and three main themes: treatment, diagnosis, and public health and quality of care (Figure 29). Majority of questions regarded treatment (81% of all questions), while others involved diagnosis and public health and quality of care (11.6% and 7.4% respectively).

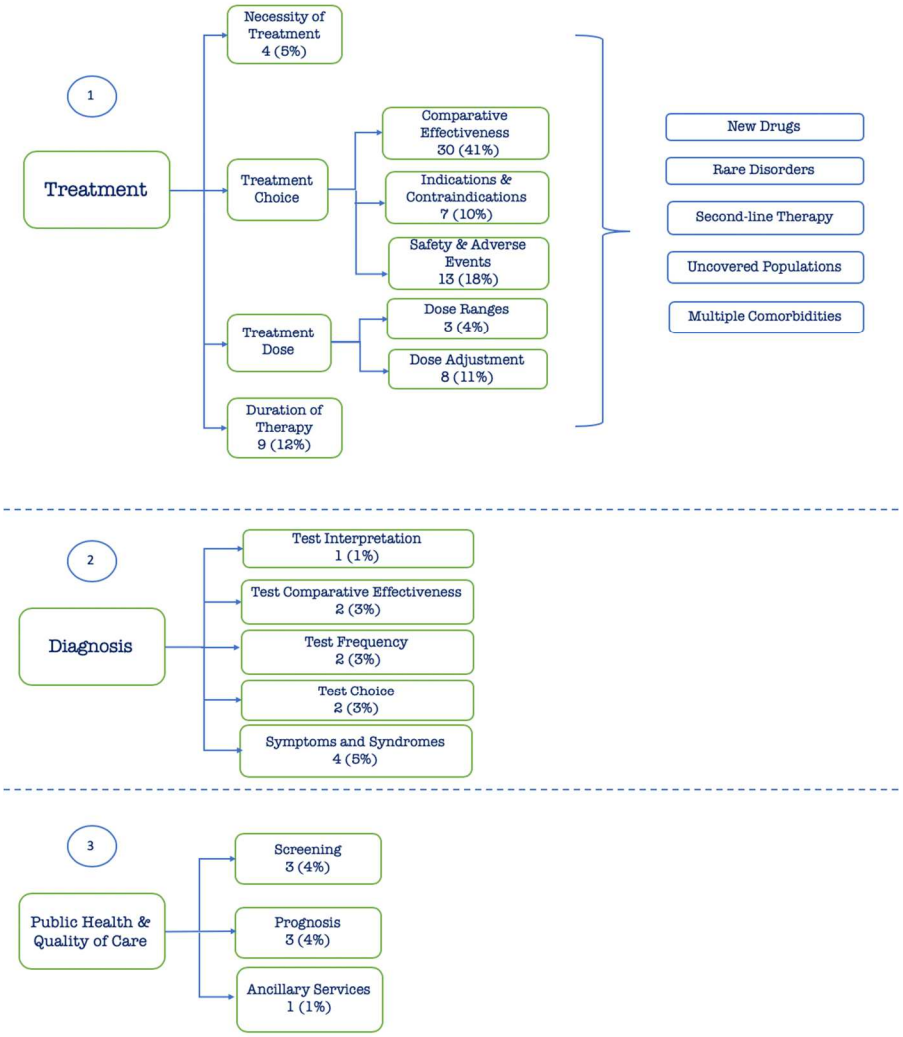


Figure 29. Taxonomy of unmet evidence-related information needs. Green boxes represent themes and subthemes and a bracket with blue boxes represents broad topics applicable to included subthemes.

Almost all the questions in the treatment group were related to drug treatment. We obtained only a few non-drug treatment questions, which did not allow us to specify this theme further. The information needs related to drug therapy further fell into four main sub-themes based on the steps in drug prescribing: decisions on drug necessity, choice of a drug and its dose and decisions on when to

discontinue treatment. Drug choice included three sub-themes and had the broadest range of possible clinical scenarios that were not covered by the existing evidence.

Clinicians reported that the current studies appeared to fail to provide evidence for newly marketed drugs ("Should a diabetic patient on ACE inhibitors, diuretics and SGLT2-inhibitors be taken off diuretics as SGLT2-inhhibitors act as diuretics?"). The available evidence also was said to inadequately cover certain populations such as pregnant women, children, elderly and patients with multiple chronic conditions and rare disorders. Such questions were usually formulated within the Population, Intervention, Comparison, Outcome, Time (PICOT) framework [89], where population corresponded to a group of patients with conditions of interest, intervention and comparison – drugs or procedures that were intended to be used for a patient of interest, outcome – to a disorder or event to be used to compare interventions and time – to time-at-risk.

Applying PICOT to the questions highlights shared common population or intervention. For example, chronic kidney disorder fell into multiple subthemes: 'Comparative effectiveness' ("Which one of the new SGLT2 inhibitors is best for patients with chronic kidney disorder?"), 'Indication and Contraindication' ("Do we know if rivaroxaban or apixaban should not be prescribed for patients with kidney failure in real practice?"), 'Drug Safety and Adverse Events' ("In patient with chronic kidney failure and hypertension taking five drugs, how do we know which one caused an adverse event?") and 'Drug Dose' ("Tacrolimus for kidney transplantation: what is the start dose, how often its level should be measured in blood?"). There was a small number of questions that do not fall into the PICOT framework (e.g., "What are the risk factors for vancomycin-induced kidney injury?") that were mainly related to characterizing patients of interest.

The ‘Diagnosis’ theme mainly included questions related to laboratory tests as well as vague symptoms and syndrome; questions related to screening, prognosis and ancillary services were classified as a broad ‘Public Health & Quality of Care’ theme.

While most of the questions were unambiguous and straightforward, others required additional clarification. For example, one clinician asked a question that at first glance could have been interpreted a diagnostic question: “How should patients be screened for dementia?”. Further discussion revealed the rationale behind the question, which allowed us to classify it into the “Screening” theme: “It [dementia] often is confused with depression especially if a person has a history of depression. We need this information to properly educate primary physicians on how to take care of such patients”.

Additional analysis

More than half of the physicians (60%) indicated an unspecified commercial tool as their primary source of evidence and only 10% turned to guidelines first to answer their questions (Appendix 5.1).

We found no correlation between the primary source of evidence and primary clinical settings (X-squared = 6.16, p = 0.29) or academic rank (X-squared = 18.67, p = 0.54). On the other hand, we found a difference in specialty and primary care practitioners: primary care physicians preferred the commercial tool (X-squared = 9.85, p = 0.08), while specialty physicians used PubMed, guidelines and the commercial tool.

We found that the number of questions related to insufficient evidence in clinical practice did not decrease with clinical experience, and in fact appeared to increase with experience ($R = 0.55$, $p < 0.01$).

Physicians who mainly work in an inpatient setting identified more information needs ($W = 120.5$, $p = 0.01$). The number of questions was unrelated to their specialty and did not change irrespective of whether they were primary care or specialty physicians ($W = 58$, $p = 0.1$).

When we analyzed the distribution of type of questions based on our themes, we found that inpatient physicians had significantly more questions related to treatment ($W = 134$, $p < 0.01$). On the other hand, the type of questions was unrelated to clinical experience or academic rank.

5.1.4 Discussion

Patterns of information needs

In this study, we found a significant gap in clinical evidence, which resulted in a large variety of information needs across different specialties and care settings. The majority (90%) of physicians in our study indicated that they practice evidence-based medicine and use evidence knowledge sources routinely. The main obstacles to applying evidence in practice were low patient compliance and socio-economic determinants along with mistrust of clinical guidelines and clinical trials. While other studies [452,455–458] listed lack of time, personal unawareness, disengagement and passivity as the strongest barriers, our findings just partially support it. Indeed, outpatient physicians could rarely find time for literature review and usually thought of their clinical scenarios as routine and straightforward. On contrary, the length of inpatient stay might give physicians more time to search for evidence. They also tend to face more challenging cases, which oftentimes requires more thoughtful research, team collaboration and experience sharing. This may imply that inpatient physicians are more likely to practice evidence-based medicine and use additional knowledge sources to find additional evidence for their clinical cases.

Surprisingly, we found that experienced physicians and tertiary care practitioners have more information needs arising from a gap in current evidence than those who have less experience. The focus of previous studies was shifted towards the most vulnerable groups: residents, family and primary care physicians [82,438,442–444,448,449,458], assuming that these categories require information the most. We have discovered that the need for clinical evidence does not diminish with physicians' maturity: while residents and practitioners at the early stages of their career [459] may experience lack of clinical expertise and knowledge, more senior physicians have already gained this expertise and oftentimes require more complicated and precise evidence. This observation relates to the Dreyfus model of Skill Acquisition [460], according to which one progresses through five levels of proficiency: novice, advanced beginner, competent, proficient and expert. At the expert level, clinicians no longer have to rely on the set of rules or analytical principles that they had been taught but use their background experience to 'intuitively' make decisions. On the other hand, experts can identify knowledge gaps and acknowledge their relevance to clinical scenarios. Constant search for evidence when no established practice exists seems to be a feature of experts, who tend to apply analytical problem solving to novel or complicated clinical cases.

We classified such clinical cases based on the main areas where physicians struggled to find answers (Figure 29). These broad areas included multiple co-morbidities, rare disorders, polypharmacy, elderly, and young patients. Our findings align with previous studies, which reported that the patient comorbidities and contexts play an important role in inability to answer clinical questions at the point of care [40,49,452].

When analyzing narrow question themes, most of the them were related to drug therapy, which reflects a persistent trend in information needs [461,462] and is well supported by the recent studies

[442,448,449]. Moreover, the pull of unanswered questions grows with an increasing number of newly marketed drugs and poor generalizability of clinical trials [43,44,463]. The distribution of the types of questions also reflects this paradigm: while thirty years ago the percent of medication-related questions was around 30% [448,461], in our study we saw that it doubled.

We observed that physicians practicing in inpatient settings were more likely to ask questions related to treatment and diagnosis, while most of the questions related to quality of care or public health-related questions were asked by the outpatient physicians. Due to its nature, outpatient care tends to be more influenced by non-clinical factors, such as socio-economic determinants or patient compliance and adherence. When prescribing treatment or diagnostic procedures, healthcare provider should account for patients' capacity to pay for their treatment and diagnosis, compliance with medications and follow-up appointments and other factors. [448,464]. For example, one of the questions ("Should we prescribe Vitamin D in children knowing that compliance may be low?") considered the relative effect of the Vitamin D tablets that are prescribed once a week. As such regimen was observed to be associated with forgotten and missed doses, the clinician required additional information on its efficacy in real-world practice. As outpatient care requires relatively fast and cheap diagnostic procedures, outpatient practitioners may be interested in evidence related to symptoms and syndromes, which would allow to interpret routine diagnostic procedures more effectively ("Can obesity inflammation alone get erythrocyte sedimentation rate up to a very high level?").

When searching for evidence to answer these questions, the commercial tool was the main knowledge source used by physicians in this study, which aligns with the previous findings [458,465]. It was said to be used for the new drugs, unfamiliar rare disorders, and disorders outside of physicians' specialty. Other physicians, especially those who preferred PubMed, characterized the commercial tool

used as a superficial knowledge base that did not provide case-specific information or in-depth knowledge ("It usually tells me what I already know. I may use it to browse new drugs, but otherwise look for new studies"). Cook et al. [458] also reported its inability to answer complex questions, which, together with our findings, may explain why specialty physicians preferred to use PubMed. The choice of primary knowledge base may also indicate the predominant characteristics of information needs arising from a gap in current evidence in specialty physicians. The latter may require evidence that addresses specific patient contexts rather than the general knowledge about a disorder and its treatment. Therefore, specialty physicians may benefit from an additional knowledge source that accounts for these needs and provides them with the evidence for a specific clinical case.

These observations support the previously observed inferential gap [43,44,463] between the guidelines and trials (which show effectiveness only for selected groups of patients) and a real patient with a complex of individual disorders, medical history and non-clinical characteristics. Not surprisingly, physicians within a specialty share the same needs and attempt to answer them based on their observations and experience. For example, we saw that some of the same questions in nephrology occurred across multiple interviews with physicians trying to answer these questions on their own. This may lead to the disparate solutions to the same clinical problem even within a single institution and, eventually, to suboptimal patient care.

Lack of evidence and inability of existing evidence to answer complex real-world clinical questions points to a need for a CDSS that can address clinicians' information needs arising from a gap in current evidence at the point of care. We can bring the pipelines and frameworks we developed in Chapters 2 and 3 to build a service ('Data Consult Service') that can generate evidence at the bedside to facilitate clinical decision-making and guide clinicians. Using EHR and administrative claims data and

domain expert knowledge, such CDSS select a group of patients similar to the patient of interest and use statistical approaches to compare it to another group that represents an alternative scenario. For example, the question "Which one of the new SGLT2 inhibitors is best for patients with chronic kidney disorder?" can be answered by conducting a comparative retrospective new-user cohort study using institutional EHR data. In such study, patients with chronic kidney disorder are classified into a target or comparator group based on their SGLT2 inhibitor exposure, balanced using propensity score adjustment and analyzed to examine relative risk of an outcome of interest. Data analysis can vary from simple descriptive statistics to comparative effectiveness studies, but generally allows one to learn from previous patient care. For example, the question "How do we interpret Synachten test in patients on long-term steroid therapy?" (Diagnosis theme) can be answered by characterizing patients on long-term steroid therapy and comparing their outcomes based on the length of therapy and their adrenocorticotropic hormone level (measured using Synachten test).

As noted in other studies [40,49,452], lack of time and difficulty remembering questions had been contributing to the pool of questions that remained unanswered by clinicians even despite the availability of CDSS and knowledge resources. Thus, a CDSS should allow obtaining relevant and concise answers in a timely manner.

Another important aspect of CDSS design is proper target group identification. Specialty physicians, compared to primary physicians, used more knowledge sources and may have more time to search for additional evidence. Specialty physicians also reported an increased need for a knowledge source that helps to answer complex context-specific clinical questions, which implies that they are more likely to benefit from such CDSS. As we observed an association between clinical experience and

the number of questions, experienced clinicians may also use the tool that provides answers to these questions.

It may seem paradoxical that we exclude novices from the target group for such tool. While novices may have more broad general information needs and may struggle with applying evidence-based practices, they would rather require guidance on applying existing knowledge to practice. On the other hand, Data Consult Service would be beneficial to clinicians who routinely apply evidence-based medicine practices, identify gaps in current evidence, and deal with complex clinical cases.

As these clinical cases are often complex and non-trivial, the team behind CDSS or the CDSS itself need to be able to identify an appropriate study design. For example, the question "How should patients be screened for dementia?" (Appendix 5.2) may be interpreted as a question about appropriate tools for dementia screening. Nevertheless, the further discussion with the clinician revealed that the question considered common misdiagnosis of dementia patients with a prior history of depression. We, therefore, would rather transform it into a characterization study of the patients with a history of depression who were and were not further diagnosed with dementia. These two groups can then be compared to identify the distinctive features of patients who had depression and developed dementia. The identified features can then be used by clinicians to promptly diagnose dementia.

Such difference in interpretations has implications on study design, the volume of EHR data needed to conduct a study and eventually on the ability to address the question. All the relevant details should be taken into consideration and a tool should allow for further question interpretation.

5.1.5 Limitations

Our study has several limitations. First, we assumed that the number of questions physicians raised during the interviews could serve as an approximation of their information needs arising from a

gap in current evidence. We believe that the time-consuming and effortful nature of literature search has made mentioned clinical cases memorable, which provides sufficient accuracy for our approximation. We did not cover all clinical specialties, but the variety of interviewed specialties allowed us to get a broad picture of information needs.

We conducted our study in a single institution, which only represents the unmet information needs in a large tertiary care hospital and does not characterize clinicians' information needs elsewhere.

5.2. Gaps in bias-mitigating strategies of evidence-generative CDSS[8]

As we observed in the previous section, physicians can benefit from a service that produces new clinical evidence at the point of care to answer the questions not covered by existing RCTs or observational studies.

Given the potential challenges associated with deploying such a service, in this section we look at the existing clinical decision support tools to learn from their experience, particularly regarding the strategies for mitigating bias and evidence quality assurance. We conduct a scoping review [466] covering a broad spectrum of CDSS that generate new evidence, from disease-specific expert-based to more complex data-driven, characterize and classify 25 tools according to their methodology and implementation stage and provide a synthesis of their features, scope and bias-mitigating strategies.

We discover that only one tool partially addresses bias, which points to the need for a CDSS with robust analytical methods and pre-analysis bias assessment.

5.2.1 Methods

Search strategy

We conducted a scoping review, including the use of a formal search strategy, appraisal of study quality, and a narrative synthesis of findings. To inform our analysis, we performed a systematic search of four databases (PubMed, Embase, Proquest, and IEE Xplore) for articles published in English before May 22, 2020.

---

[8] This section is published in JAMIA. The full citation for this publication is:

Ostropolets A, Zhang L, Hripcsak G. A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. *JAMIA* (2020*)*

We identified four components of our search: (1) electronic health records, (2) clinical decision support tools, (3) evidence-based medicine and (4) complex clinical cases.

For each component, we included the hyponyms, synonyms and broader terms identified in the previous literature and through discussions with librarians [443,452,461,467–470]. The search terms included MeSH concepts (PubMed), EMtree (Embase), and free-text terms, combining the four groups of terms described above (Appendix 5.3).

We included articles that described any type of clinical decision support system, which was defined as any computer system designed to impact clinician decision making about individual patients at the point in time that these decisions are made [471], that use patient data to address clinical questions not covered by existing evidence and are designed to be used by clinicians. We focused on those CDSS that either modify existing evidence tailoring it to the patient of interest or generate previously unknown knowledge to facilitate decision-making.

We excluded articles meeting any of the following criteria:

1. The CDSS only used existing evidence (clinical trials, guidelines, published literature)
2. The study was in a language other than English
3. The study used data other than clinical (for example, genomic or protein data) or simulated datasets

Review articles were used to obtain relevant references and to inform this discussion.

Study selection and data synthesis

Two reviewers (including AO) independently screened the title and abstract for each study for inclusion and exclusion criteria. The level of agreement between the two reviewers was assessed by a

Cohen's kappa score. Disagreements between the two reviewers were resolved by discussion until consensus was reached. The list of studies selected for full text review was screened for relevant references. AO reviewed the full text of the selected studies and extracted the year of implementation or evaluation, the site of intervention, its main area (specialty), focus (patient care, research, quality improvement), methods used, evaluation type and evaluation outcomes. Extracted data was reviewed and approved by the second reviewer.

5.2.2 Results

We retrieved 3427 articles, out of which 172 articles were potentially relevant based on the abstract, title, and keywords screening. 144 papers were identified as duplicates and removed. The level of agreement between the two reviewers was reflected by a Cohen's kappa of 0.84.We additionally found 83 articles through reference lists. 53 manuscripts describing 25 CDSS were selected for this review (Figure 30).

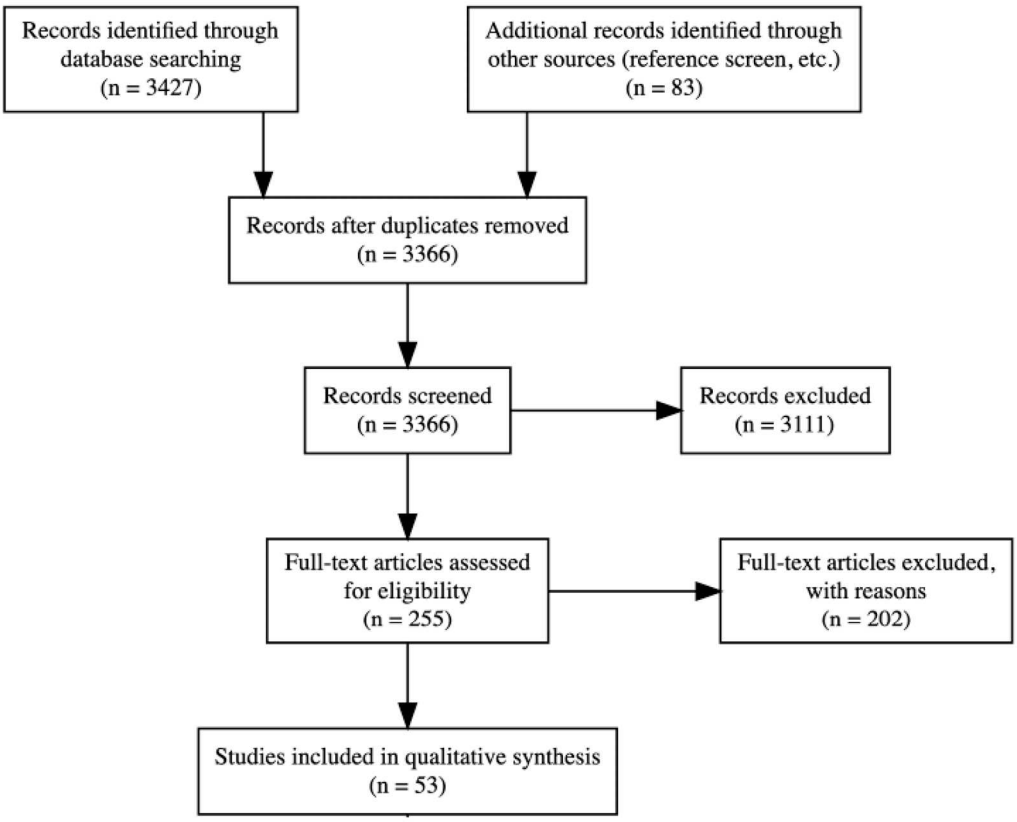


Figure 30. PRISMA flow chart describing the article selection and review for the scoping review of the clinical decision support tools that generate new evidence.

Only 36% (n=9) of the tools were implemented in more than one site with 64% (n=16) of the tools developed in the USA, the others – in Europe (6), UK (1) or China (2). All CDSS included in the study were either used or planned to be used for clinical care and 68% of CDSS also focused on quality of care or research (eight (32%) and nine (36%) respectively, Appendix 5.4). Oncology was the main area of use (nine CDSS, 36%) followed by surgery, psychiatry and internal medicine (two, 8% each). The other tools did not have a specific area of use, albeit specific use cases were used to show the

features of some prototypes (potentially unrestricted). Only 40% (n=10) of the tools were deployed and used in practice.

We classified the tools into two groups based on the main approach used to infer new knowledge: (1) data-driven tools, which use patient data to generate practice-based evidence [472] in real time and (2) expert-based tools, which require experts to incorporate practice-based evidence into algorithms subsequently used in CDSS. Both groups produce knowledge that does not explicitly exist outside of CDSS and should be useful for decision-making for a patient of interest. Based on the analytical component, data-driven tools can further be classified into (1) visual non-analytics-based tools and (2) analytics tools (Table 14).

Table 14. Included articles grouped by the inference mechanism and analytical component.

Group	Sub-group	Included CDSS
Data-driven CDSS (n = 17)	Visual non-analytics-based tools (n = 6)	"Composer", "ePEPS", "CaVa", "CareFlow", "Patient-like-mine", "PatternFinder"
	Analytics tools (n = 11)	"Care Pathway Workbench", "Green Button", "CoCo", "VisualDecisionLinc", "Melanoma Rapid Learning Utility", "DICON", CDSS for radiologists (Morrison et al.), two CDSS for prostate cancer (Bernard et al.), two CDSS for diabetes mellitus and acute coronary syndrome (Xia et al.)

Expert-based CDSS (n = 8)	"MayoExpert", "e-bipolar", ROAD2H CDSS, "Oncology Expert Advisor", "P4 Pathways", "Level I Pathways", "ViaOncology", "eviti"

Data-driven tools

Visual non-analytics-based tools

Visual non-analytics-based tools allow defining patients based on the criteria of interest, aggregating them according to a set of rules defined by a clinician and visually inspecting resulting patient cohorts. Individual patient data or aggregated data can be aligned by timeline and presented to a clinician for comparison. With rare exceptions [473] they do not require a third party to perform analysis, so that the users can obtain relevant information on their own. Patient criteria can be selected from a pre-defined list [473,474] or from any structured data in the EHR system [475]. The latter increases the variety of questions clinician can ask as any diagnosis, procedure or laboratory test from the EHR system can be selected. While such tools are not capable of generating gold-standard evidence, they allow clinicians to learn from previous care, observe and compare patient outcomes.

One of the tools, "PatternFinder" [476–480] visualizes patient records according to temporal queries and allows specifying an index event and two additional events only. Visualization is limited to matched events, so that clinicians can only explore the common events but not those that differ among patients. The authors performed extensive 4-months usability testing for patients with contrast nephropathy.

"CaVa" [473] visually represents the changes in the pre-defined variables of interest through line thickness, where lines connect clinical events and are aligned by timeline. As opposed to "PatternFinder", it does not limit the number of events to display but requires the study team to identify patients and select variables in advance. A prototype, developed for cardiology patients, could also support similarity measurements and utilization analysis. "CareFlow" [481], developed independently, has a similar interface and features. It was similarly tested for cardiology patients (congestive heart failure).

A CDSS for radiologists presented by Morrison et al. [482] aims at identifying patients with similar demographic and lung cancer-related characteristics. It displays descriptive statistics mainly focusing on cancer characteristics and unlike the other tools it is characterized by limited visualization. Another distinctive feature is repurposing of a lung screening trial data set, which limits its ability to learn from the new patient data.

Similarly to the previous tool, "Composer" [474] was developed for a single specialty, assisting orthopedic surgeons in assessing patient state after spinal procedures. The developers pre-specified outcome measures and subsequently plotted them for individual patients or aggregated cohorts.

"ePEPS" toolbox [475,483,484] was built on top of the nationwide French EHR database, leveraging the benefits of linking patient data across multiple institutions. It supports constructing patient cohorts based on all available structured data, not limiting variables to a pre-defined set. Clinicians can then compare the cohorts based on geographic distribution, explore the distribution of the events of interest across groups, and inspect patient trajectories.

A tool presented by Li et al. [485] used elastic search to search for patients of interest in real time. The tool preserves the ability to transparently explore "what-if" scenarios for cohorts of patients. It

is achieved by comparing the trajectory of a given person to the trajectories of similar patients and visually analyzing if his or her trajectory is within normal bounds.

Analytics tools

Analytics tools aggregate patient data and use statistical approaches to compare patient cohorts. Data analysis can vary from simple descriptive statistics to comparative effectiveness studies, but generally allows not only to learn from previous patient care but also to reliably compare patient outcomes and characteristics. The tools described in this section mainly differ in the methods used to obtain cohorts and the ways results are presented to clinicians.

We start with the tools that combine visual representation of patient cohorts with statistical analysis that allows clinicians to obtain estimates (odds ratios or relative risk) for the groups of interest.

"CoCo" [486], which stands for "Cohort Comparison", provides visualization of interactively refined patient cohorts as well as individual patient records as time sequences. Clinicians can then compare the cohorts using formal statistical approaches (survival analysis and log-rank test).

Bernard et al. [487] developed a dashboard similar to "CoCo" to visualize multiple patient trajectories for patients with prostate cancer. Clinicians can select patients and compute correlations between variables and a patient cohort. Patient histories are synchronized, and each history is shown as a line with a color corresponding to a phase of prostate cancer treatment. The authors further explored this area by developing dashboard networks to allow cohort comparison [488]. To conduct user acceptance testing Barnard et al. asked clinicians to run an observational study using the prostate cancer dashboard. They showed that the system increased the efficiency of analytics and provided visual assistance for complicated temporal relationships in the data. Malik et al. [486] took the same evaluation approach for

"CoCo" and generated use cases to access perceived usefulness of the tool in emergency department settings.

"Melanoma Rapid Learning Utility" (MLRU) [489] and "VisualDecisionLinc" [490] similarly enable physician-driven cohort selection and comparative effectiveness analysis for melanoma and major depressive disorder respectively. Physicians construct cohorts based on demographic data, drug exposure data and melanoma-related variables (MLRU) and can subsequently inspect odds ratios of outcomes produced by survival analysis. MLRU underwent user acceptance testing with positive feedback and physicians more interested in using it for research purposes. It was deployed, but the actual use of the tool has not been reported.

Nan Cao et al. used a glyph-based visualization system (DICON [491]) to show structured data from patients' EHR and compute similarity scores across patient data elements. The tool calculates the correlation between selected features (e.g., International Classification of Diseases, 10th Revision, Clinical Modification codes) and the cluster of similar patients. The authors conducted a formal mixed (qualitative and quantitative) user acceptance testing with domain experts and non-experts focusing on the design of the icons that represented patient cohorts. They found that their design provided higher efficiency on group comparison; they provided limited information about user feedback.

Xia at al. [492,493] developed two separate prototypes that use clustering techniques to identify patient sub-groups within a disorder (acute coronary syndrome and diabetes mellitus). Upon clinical encounter, a clinician can see a cluster of patients who are similar to the patient of interest and inspect their characteristics, including demographics, disease onset and progression, drug exposure and outcomes.

As opposed to visual analytics-based tools, “Green Button” [494–496] presents cohort comparison in the form of reports that the developers supply to clinicians. Developed primarily for clinical care, it leverages observational studies to answer clinical questions ranging from simple descriptive statistics to comparative effectiveness studies. A fast search engine retrieves patients of interest and visualizes their medical events as temporal sequences, allowing fast and efficient iterations. This is the only tools that mitigates confounding (propensity score matching) in addition to survival analysis, incidence rates or descriptive statistics. Compared to the other tools described here, clinicians have to supply their questions to the study team and cannot execute analysis on their own. The CDSS delivers reports through protected e-mail and phone conversations rather through a standalone user interface.

The last tool in this section, “Care Pathway Workbench” [497] transitions from pure data-driven approaches to integration of newly generated real practice evidence into existing care pathways. It uses Hidden Markov Models to identify the deviations of real practice from clinical guidelines and mines EHR data to obtain clinical event sequences. It then presents these insights from real-world practice to clinicians so that they can modify care plans for a specific patient.

Expert-based (knowledge-aggregative) tools

Expert-based CDSS rely on a study team (usually a multidisciplinary clinical team) to synthesize multiple sources of evidence into a knowledge base incorporating evidence-based recommendations and local insights of previous patient care and outcomes. Similarly to the traditional rule-based CDSS, such tools use an existing evidence knowledge base complemented by the newly generated practice-based evidence, which is not available outside of the tool.

A significant portion of the tools presented in this group relates to cancer care. The latter is characterized by multiple "best" treatments that accommodate specific patients' characteristics. Such treatments or pathways are often selections of the most cost-effective treatments and are developed collaboratively with local specialists. "P4 Pathways" [498–500], "ViaOncology" [501,502], "Level I Pathways" [503–509] and "eviti" [510,511] have been widely adopted across the United States. They aggregate and modify cancer clinical guidelines according to the community-based practices and practice-based evidence so that these human-curated pathways reflect the way care is delivered. These tools can be used to query EHR data and obtain the cohorts of similar patients for whom specific pathways are applicable.

Disadvantages of such tools include sensitivity to incomplete data, dependence on local experts and previous practices, and focus on treatment cost. On the other hand, they have established feedback loops for fast evidence modification upon practices changes, or new knowledge becomes available. As treatment pathways are curated by the leaders in the field, they can be perceived as trustworthy, which may have facilitated their adoption [512].

Pathway-related tools underwent extensive evaluation, including patient and quality outcomes. "Level I Pathways" was made available to the clinicians within the US Oncology Network (a network of more than 400 integrated, community-based oncology practices) in 2013 and was proven to reduce costs of treatment for patients with lung, breast and colorectal cancer; reduce duration of treatment; and lower cancer-related re-admission rates [503,505]. Nevertheless, no statistically significant difference in survival rates was found. "P4 Pathways" reduced inpatient admission rate and duration of therapy for chronic lymphocytic leukemia [499]. "ViaOncology" showed the same results for metastatic colorectal cancer [502].

"Oncology Expert Advisor" [513] is a closely related CDSS that also provides pathway-like recommendations related to cancer care. While it also aggregates multiple sources of evidence, the core function of this tool is to promote sharing best practices by incorporating peer-to-peer consultations based on the patient profile created by this tool. It subsequently includes the advice management system that allows consultation tracking.

ROAD2H [514,515] and MayoExpert [516] are two other practice-based evidence learning health systems, which aggregate recommendations from international, national, local guidelines and institutional practices to provide tailored knowledge. ROAD2H uses argumentation with a clear provenance trail to resolve conflicting recommendations, while MayoExpert represents care models as sequence of nodes, where a node is a decision point. They provide clinicians with a patient-specific recommendation based on hospital EHR data prioritizing institutional best practices. ROAD2H has been piloted in two sites and currently provides recommendations for patients with chronic obstructive pulmonary disorder and chronic kidney failure. MayoExpert, on the other hand, incorporated 106 models at the time of publication at the time of publication and was used by 60% of clinicians at the Mayo Clinics sites. The authors found that general practice specialists and less experienced practitioners used the tool more often than specialists and more experienced clinicians.

"e-bipolar" [517] stands on its own in this review. As opposed to the "top-down" approach used by pathway-related tools, "e-bipolar" helps French practices in getting practice-based advice from the other specialists. Coordinating center manages assessments, provides guidance on optimal treatment and shares practice-based evidence by providing anonymized data through a web-application "e-bipolar".

5.2.3 Discussion

In this paper we explored the tools that aim at guiding clinicians in complicated clinical cases for which they do not have gold-standard evidence. Existing reviews focus on the tools that facilitate evidence-based practices, but the latter cannot answer all questions outside of guidelines or trials [518–520]. Meanwhile, the availability of knowledge plays an important role in the quality of decision-making (Appendix 5.5) [521–523]. For questions not covered by existing evidence, clinicians must rely on their limited experience. For example, there is no clear consensus on common clinical questions like "Should a diabetic patient on angiotensin-converting enzyme (ACE) inhibitors, diuretics and sodium glucose co-transporter 2 (SGLT2) inhibitors be taken off diuretics as SGLT2 inhibitors act as diuretics?", which results in clinical practice variation.

Among other solutions, CDSS can generate additional knowledge to guide clinicians. Visual tools focusing at presenting longitudinal patient data has been known for a long time, starting with LifeLines [524] and KNAVE [525]. They evolved in two directions: (1) adding more sophisticated features to individual views and (2) aggregating patient data into groups with a subsequent visual or statistical analysis. The main highlights of visual systems are automated process, fast execution, flexibility, relatively small maintenance cost and intuitive representation of the results. On one hand, an ability to quickly explore aggregated patient data facilitates fast answers to clinical questions in real time. It also means that tool utilization is relatively cheap as it does not rely on a team supporting query execution and report generation. On the other hand, it demands familiarity with the data, which can be unfeasible for non-experienced clinicians. Additionally, such tools are inferior to analytic CDSS in terms of scientific rigor as they do not imply that observed differences in patient cohorts are statistically significant or unbiased [67].

Another approach, implemented in “CaVa”, “Green Button” and expert-based CDSS, requires a third party (study team or experts) to generate knowledge either by running small-scale observational studies on patient data or incorporating new practice-based evidence into personalized recommendations. An advantage of such approach is involvement of skilled professionals, who are familiar with the data and research methods. In expert-based tools, the knowledge is gathered in advance and then tailored to patient’s characteristics. New knowledge, therefore, cannot readily be made available if complicated clinical scenarios are not covered by the existing pool of care models. As timely answers can be critical in decision-making, another approach adopted by “Green Button” is to run observational studies in real time. While it can address a broader spectrum of questions in a timely manner, such a type of CDSS must rely on efficient communication with clinicians to capture additional details and refine questions. On the other hand, such CDSS have a potentially broader audience since they do not require specific skills or knowledge and the results can be interpreted by skilled personnel or introduced in a simplified form in the reports.

Regardless of knowledge inference methods, CDSS for new evidence generation were mainly developed and implemented at one site and rarely disseminated. Data-driven tools mostly used structured data (ICD-10, CPT-4 and ATC codes); unstructured data was processed by only two tools [485,526,527].

None of the tools harmonized data from disparate data sources or used common data models [528]. Lack of data standardization can pose a challenge if generating knowledge requires gathering data from multiple sites or sources, for example, if a disorder or event is rare.

Lack of evaluation is another finding in our study. For some of the tools, there was no information about evaluation including the types of tests that should be performed at the initial stages. The pathway-related group was the only group for which an impact on patient and quality outcomes has been shown. These CDSS are based on pathways supported by payors [498], which may be a reason for their wide adoption and evaluation. Another possible reason may be expert involvement.

As long as traditional RCTs cannot deliver sufficient evidence on time, such tools may be a good alternative to disparate intuitive clinical practices. Due to the limitations of the current tools, new robust CDSS may be needed. They should build upon previous designs and incorporate their strength in delivering new evidence at the point of care. Most importantly, they should implement more robust strategies to ensure reliability of evidence.

Ease of use and intuitive result presentation should be combined with robust statistical methods and phenotyping. While full-scale observational studies usually undergo rigorous assessment [56,83], small-scale real-time studies may not produce unbiased estimates. For example, rule-based phenotyping with chart review validation [114] may not be feasible in real time, which creates a need for best practices for fast yet accurate patient identification. If a tool aims to answer questions from different areas or specialties, the ability of a particular data source to supply accurate data should be articulated to clinicians and any data quality issues or other limitations should be acknowledged. If phenotyping is done by an individual other than the end user, phenotyping principles, accuracy and limitations should be transparently described as well. Regardless of the design used, a CDSS has also to be seamlessly integrated in the workflow.

5.3. Generating and delivering evidence at the bedside: Data Consult Service [9]

Based on the assessment of clinicians' information needs and review of existing clinical decision support tools, we proceed with designing a service (Data Consult Service) for generating new evidence at the point of care, which specifically focuses on providing robust and reliable evidence.

We run a pilot study at NewYork-Presbyterian hospital, which results in answering 24 questions from 22 clinicians, and start a knowledge base that describes current challenges related to use of observational data in real time evidence generation.

We show that such a service is capable of supplying new evidence to clinicians to inform their decision-making and partially address their information needs. In providing such a service, it is mandatory to ensure reliability of delivered evidence by accurately phenotyping patients of interest, assessing the quality and completeness of the data, and using appropriate research methods to mitigate bias.

5.3.1 Background

Despite the growing body of medical knowledge, a substantial number of clinical questions remain unanswered [13,43,49–52]. Observational data can be used to address some of them with publications and presentations being a common way to disseminate new evidence. Nevertheless, observational studies are still time-consuming with an average study taking up to year [72,260].

[9] This section is published in JAMIA. The full citation for these publication is:
Ostropolets A, Zachariah P, Ryan P, Chen R, Hripcsak G. Data Consult Service: Can we use observational data to address immediate clinical needs? *JAMIA* (2021).

As we showed in the previous section, new CDSSs ranging from visualization tools to complex learning systems aim at generating new evidence in real or near-real time [14]. While promising, such CDSSs are generally not scalable as they usually focus at one condition or area at a time and are oftentimes descriptive in nature as they do not address bias and confounding properly.

Given known limitations and pitfalls of observational data [84], it is unclear to what extent observational data used by such tools can address clinicians' immediate information needs [496]. It is unclear if the methods used to mitigate bias can be applied in a timely manner to ensure the quality of evidence generated at the point of care. There is, therefore, a need to identify the scope of the immediate clinical information needs observational data can address and, more importantly, the pitfalls that have to be considered.

Second, there is limited knowledge on the use of this group of tools in real clinical practice. Most of the tools that were deployed in clinical settings and showed improved outcomes involved traditional rule-based approaches [14]. On the other hand, data-driven CDSSs remain limited to a single center and are rarely used.

Similar to the Green Button project [496], we launched a pilot project called the Data Consult Service that uses observational data to produce new knowledge and facilitate clinical decision-making in near- real time.

5.3.2 Methods

We launched a pilot study of the Data Consult Service with the clinicians affiliated with Columbia University Irving Medical Center aiming at assessing the feasibility of the project and the ability of observational data to meet clinicians' needs. We designed and implemented a pipeline, which involves five steps (Figure 31) starting with clinician recruitment and question gathering.

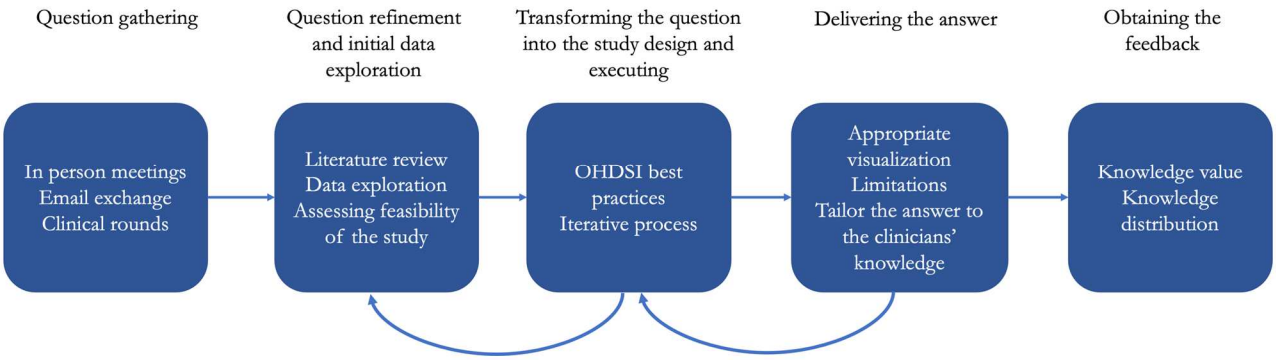


Figure 31. Data Consult Service pipeline.

Clinician recruitment and question gathering

Initial recruitment of clinicians affiliated with CUIMC was done using snowball sampling strategy [529] through email communication and in-person meetings. Clinical questions were subsequently collected at the initial or follow-up encounter through email communication, in-person meetings or clinical rounds, whichever was more convenient for the clinicians. We collected routine clinical questions that can be answered with aggregated patient data, such as questions related to practice patterns, treatment pathways, patient outcomes and others. We did not provide identifiable patient-level information. During March – December 2020, the consults have been limited to email communication due to the restrictions placed on in-person meetings.

Question refinement and initial data exploration

After clinicians submitted questions, we clarified them and reformatted according to the Population, Intervention, Comparison, Outcome, Time (PICOT) framework [89]. As our study team

designed and executed studies, we assumed that our target users had little experience with data processing or research methods. Therefore, we additionally clarified the rationale behind the question to apply an appropriate study design (incidence or prevalence rates, treatment pathways, comparative effectiveness, predictive analytics) and asked for known issues related to data capture in local EHR system (for example, known confounders or missing data elements).

After the research question was fully formulated, we proceeded with initial data exploration to assess the feasibility of the study. It included identifying necessary data components, estimating sample size and accessing data plausibility. Data exploration and analysis relied heavily on the OHDSI infrastructure: OMOP CDM, Standardized Vocabularies, tools and pipelines. If a question was deemed to be addressable, we created phenotype algorithms for identifying patients of interest. Such algorithms were written in SQL or using the OHDSI tool Atlas, which was used to create cohorts of patients by defining inclusion and exclusion criteria using available structured data. After defining an initial set of patients, we explored patient characteristics (demographic information, incidence rates, co-morbidities and other relevant information). We iteratively refined the definition and randomly reviewed individual patient histories to ensure that the patients represented the target population of interest, assess missing data and plausibility of patient profiles.

Observational study execution

After defining cohorts of patients, we proceeded with running the observational study, which was designed in Atlas or SQL and executed in R. The study design spectrum included patient characterization, treatment pathways and incidence rate analysis, population-level effect estimation (using comparative cohort, case-control, self-controlled case series, self-controlled cohort, or case-crossover designs) and patient-level prediction studies. The OHDSI infrastructure provided a seamless

study execution environment and ensured that validated observational research practices were used. The OHDSI observational research framework [74] emphasizes a systematic process for generating reliable evidence, such as pre-specifying the study design any data analysis to avoid P-hacking, mandatory application of methods to control confounding (large-scale propensity score matching using all available covariates [134,530]) and examination of study diagnostics including empirical evaluation through the use of 15-100 positive and negative controls [204] to detect residual bias and for p-value and confidence-interval calibration [531]). Such practices increased the transparency of analyses and allowed the study team to assess potential biases and decide if the results should be delivered to clinicians.

After a full study specification was generated, we selected appropriate data sources for the study based on the target population of interest, necessary data elements and available patient sample size. The list of data sources included CUIMC EHR, MDCD, MDCR and CCAE (Table 1). If applicable, we ran the study in multiple data sources to examine the consistency of findings. Each database had been transformed into the OHDSI OMOP CDM version 5 and had been used in numerous studies [24,195,215,259,532–535]. Additionally, we leveraged the multi-step quality assurance process adopted by the OMOP CDM [100,101,536], which comprises checks for data plausibility, conformance and completeness. Upon study result generation, we examined the output for potential bias (e.g. propensity score balance and comparison of calibrated and non-calibrated estimates) as well as overall plausibility (via review by two team member physicians). The list of potential biases we assess include participant selection bias, attrition and detection biases, confounding and reporting bias [215,533,537,538].

Answer delivery

We compiled the results into a study report, which contained a summary of the study following STROBE guidelines [539]: study design, our findings and appropriate visualizations (plots, charts etc.), study limitations and potential biases, along with the information about data sources used. We used visualization techniques to tailor the reports to clinicians' knowledge of research methods and data. If requested, we performed additional post hoc analyses, such as studying other patient subgroups or conducting additional as-treated or intent-to-treat analyses.

Upon delivering the report, we discussed the results and their limitations with clinicians and collected users' feedback. We asked if the report answered their question, if it was comprehensive and easy to understand, if it aligned with their experience or prior beliefs and if it was likely to change their practice.

5.2.3 Results

Overall results

At the time of writing this manuscript, we have collected 29 research questions from twenty-two clinicians, with 24 (83%) having been answered or in progress (Appendix 5.6). The other five (18%) questions could not be answered due to a lack of data elements. While most of the clinicians (19 clinicians, 86%) supplied one question, the others supplied up to seven questions. We observed that clinical rounds gathered more questions per person as we collected questions during clinical care, while email communication and in-person meeting were associated with fewer questions asked. Also, as our service had already been in place during Spring 2020, COVID-19 related questions could have only been supplied by email due to the clinical service disruption.

As we started our initial recruitment among internal medicine specialists (cardiology, infectious disease and pediatrics), the questions were mainly related to infectious disorders (9, 31%), cardiology (8, 28%), nephrology (6, 21%) and COVID-19 infection (5, 17%).

We also classified questions based on what the answers are intended to be used for (Appendix 5.6, characteristics). Most of the questions (17, 59%) considered a group of patients, for which those questions recurrently emerged over time. For example, the question "What is the relative risk of major cardiovascular events and bleeding within two years after anticoagulation therapy initiation in patients with end stage renal disorder treated with warfarin compared to patient treated with apixaban, rivaroxaban or dabigatran?" was relevant to a large group of patients and was requested by multiple specialists. Six (21%) questions (predominantly collected during clinical rounds) considered a specific patient treated at that time. For example, the question "How often Kocuria marina can be seen in the microbial culture?" is unlikely to be highly relevant for other patients with infectious diseases due to rarity of this bacterium. Finally, the other six primarily considered research questions intended to be published and, in this way, to be influencing clinical decision-making for a larger healthcare audience.

User feedback

The Data Consult Service received positive feedback with most of the users willing to share new knowledge (Appendix 5.6) with their peers. For all questions that we answered, the clinicians reported that the service met their information needs. For 8 questions, they also expressed interest in further research. Answers mostly aligned with prior clinicians' beliefs and did not require changes in current practice patterns. For these questions which aligned (20 out of 22 answered, 91%), the clinicians indicated that would use the results in their practice and disseminate the findings. The results for the

other two (9%) questions did not align with prior clinicians' beliefs, and they stated they would not change practice based on the study results.

All clinicians commented that the reports were easy to understand, even when the research methods used did not align with original study design formulated by the clinicians. As reports underwent iterative changes based on user feedback, all the users were satisfied with the reports' quality and comprehensiveness. Additional comments provided in Appendix 5.6 mainly concerned the ability of observational data to capture patients and events of interest, including underreported conditions or lack of data elements in structured data.

Questions characteristics

A large portion of questions (17, 58.6%) were answered using incidence rate or patient characterization design, followed by drug comparative effectiveness and safety (12, 41.3%). Patient characterization included computing descriptive statistics in patient groups of interest including summarizing key features and estimating incidence rates of outcomes. We also included characterization of treatment pathways (as in question 13, Appendix 5.6) in this group. Most of the questions collected on clinical rounds involved incidence rate or patient characterization design, while questions gathered through indirect communication mainly involved comparative effectiveness study design. Question complexity varied greatly. Comparative effectiveness studies were on average more complex as hypothesis testing involved methods for asserting causality and mitigating bias.

Processing time varied greatly with comparative effectiveness studies taking up to a week to produce reports. Incidence rates and characterization questions were usually answered within a day with up to five days needed to discuss the results with clinicians and adjust the reports to their knowledge. While constructing reports, we used a template that included original questions, methods used and the

description of the data source. Nevertheless, the process of writing and tailoring reports appeared to be the most time-consuming part of the pipeline. On average, a report consisted of four pages (examples provided in Appendix 5.7) and took up to two days with additional modification to clinicians' needs. Because the questions were aimed at addressing questions in clinical care and not purely research, output was formatted to inform the clinical requestor and support decision-making, rather than framed as a publication. Nonetheless, one study was of sufficient interest to reframe and expand into a clinical journal article [17].

Use of observational data for real-time evidence generation

Both comparative effectiveness studies and patient characterization required accurate patient phenotyping, which appeared to be the main category of issues we encountered. Accurate phenotyping was infeasible for some of the questions due to the lack of data elements (Table 15, "Events poorly captured in structured data"). For questions related to drug therapy in patients with chronic conditions, drug adherence, therapy modification and over-the-counter drug therapy were the main issues.

For example, when attempting to estimate a relative risk of arrythmias in patients on different anti-diabetic drugs (question 8, Appendix 5.6) we had to design cohorts accounting for the fact that patients oftentimes change their antidiabetic therapy and can be on multiple drugs simultaneously. For the same question, the drug exposures were recorded as pharmacy prescriptions which did not necessarily imply that the patients took those medications. Additionally, the prevalence of ventricular fibrillation (one of the outcomes) in the EHR was lower than the average prevalence in the population, which suggested insufficient capture of this disorder in structured data.

Table 15. Groups of data-related issues observed when designing, conducting and reporting studies in Data Consult Service arranged around measurement error and pre-analysis bias.

Group	Examples
Measurement error	
Duration of drug therapy identification	How to properly identify possible exposure gap for dual antiplatelet therapy to be considered continuous when multiple duration of therapy exists in practice?
Over the counter drugs	Information about famotidine exposure may be missing in the EHR database as it is an over-the-counter drug.
Underreported conditions (events poorly captured in structured data)	Non-life-threatening allergic reactions are generally poorly recorded within EHR, which requires additional clinical note analysis when attempting to answer questions related to drug allergy; Preliminary data analysis revealed unusually low prevalence of deep venous thrombosis.
Non-specific coding	In-stent thrombosis is coded as non-specific code (Other complications due to other cardiac device, implant, and graft), which obstructs proper patient phenotyping.
Pre-analysis bias	

Identifying appropriate study design	Given non-random fashion of COVID-19 testing and therapy, a COVID-19 related study has to be carefully designed to mitigate bias.
Drug adherence in outpatient prescriptions	Actual drug exposure is unknown in patients on oral anticoagulants as pharmacy prescription filling is recorder in the EHR database only for a subset of patients.
Frequent therapy modification	In antidiabetic drug comparison, how to identify target and comparator groups given that patients often switch therapy or stay on multiple antidiabetic drugs?
Other issues	
Limited sample size	There is a limited number of patients with breast cancer on estrogen receptor blocker or aromatase inhibitors who underwent COVID-19 testing in the EHR database.
Noncompatible groups (study diagnostics failure)	Study diagnostics reveals unsatisfactory propensity score balance when comparing patients on warfarin and direct oral anticoagulants
Data missingness	As ceftriaxone is believed to increase bilirubin levels in neonates, we expect such patients to have bilirubin measured. How should we interpret a large number of patients without bilirubin measurements?

On the other hand, questions that involved prevalent conditions, clear principles of coding or low treatment variation were overall less time and effort demanding.

For a portion of COVID-19 related questions, preliminary results were generated but would require larger sample size to produce reliable estimates; a subset of questions were not answered due to a lack of data elements. The latter involved data elements not converted to structured data at the time of analysis (echocardiography or blood culture reports) or data elements of unsatisfactory quality (vasopressor infusion regimens).

5.3.3 Discussion

Generating evidence at the point of care enables reliably answering clinical questions that otherwise remain unanswered until evidence from clinical trials or observational studies is published. As previously shown [540], a lack of unambiguous evidence may contribute to variability of clinical practice and lead to suboptimal patient care [38]. As shown in our pilot study, the Data Consult Service enables real-time evidence generation for such questions, both those that recur over time and those that are only applicable to individual patients. While the latter are proposed to be addressed by multiple precision medicine initiatives [38,541,542], we also address the questions related to a larger group of patients, which are not covered by the existing evidence. Although recurrent clinical questions point to a need for formal research projects to disseminate evidence, the Data Consult Service did still fill the need in a timely manner. It showed a potential to meet clinicians' information needs and provide new evidence that is likely to be disseminated within a healthcare institution.

Despite its advantages, generating evidence on time poses multiple challenges. Clinicians have to identify gaps in their knowledge and communicate them to the study team. As previously shown, lack of time and complicated access to information resources prevent clinicians from pursuing their questions

[40,49,50]. In this study, we found that interacting with clinicians during clinical rounds were the most productive and convenient way to obtain clinical requests. Similar findings have been shown in the clinical librarianship program, which highlighted direct presence of librarians in clinical settings to tailor answers to questions to specific clinical context [543]. Questions can be gathered in real time through rounds and do not require additional effort to submit to the Data Consult Service team. Clinical rounds ensured seamless transition of questions from clinicians to the study team and minimized the risk of being forgotten. Moreover, being present on rounds allowed us to participate in the discussion and support clinicians in recognizing potential needs. Compared to other "on-demand" services [494], Data Consult Service included participation in clinical rounds, which allowed pro-active collection of clinical questions.

Measurement error and patient phenotyping

One must ensure that the evidence is reliable. Accurate patient phenotyping and assessing data quality are critical in producing recommendations for clinicians. Previous studies [494,496,544] emphasized a need for a fast search engine, which allows quick iterations on patient cohorts. Our experience shows that search time is not the main constraint in phenotyping as opposed to having reliable approaches to phenotyping. The latter oftentimes requires advanced data exploration, which goes beyond identifying people with ICD-9/10(CM) condition codes, which are used in other studies [14]. Missing or inaccurate data in observational data sources may obscure valid inference, which makes it important to identify possible pitfalls and biases prior to informing clinicians. Developing a phenotype library to improve re-use of previously applied definitions and establishing a standardized framework to design and evaluate phenotypes could greatly improve the quality and efficiency of future Data Consult Service activities.

As a large portion of questions considers drug exposures, phenotyping oftentimes requires accurate identification of exposed patients, including combination therapy, patient switching and discontinuing drugs and over-the-counter drugs. As previously noted, the estimation of drug adherence [296] may be complicated when prescription filling is recorded for only a subset of patients. We encountered similar issues and generally considered inpatient drug administration as more reliable. Nevertheless, we had more confidence in outpatient prescriptions if patients had reoccurring prescriptions every three or six months, which indicated that they were likely adherent to a treatment regimen. Additionally, over-the-counter drug exposures were rarely captured in the EHR, which required phenotype algorithm modifications whenever such drugs were a part of phenotype of treatment.

Similarly to over-the-counter drugs, some of the studied disorders were underrepresented in the structured data. Although this trend has been shown in administrative claims datasets for acute [545] and chronic [9,273] kidney failure, thromboembolism [546], ventricular arrythmias and cardiovascular death [547] there is less research for EHR sources. The clinical informatics community would benefit from a comprehensive list of such disorders as underreporting may have direct implications on study results, especially if misclassification bias is differential [69].

Also, a portion of questions could not be answered due to the lack of data elements present only in clinical notes or reports. As opposed to the category of abovementioned conditions, such elements (for example, echocardiography data) are generally present in observational data but are missing in a particular instance. These elements can potentially be added to our OMOP CDM instance in the future by applying natural language processing techniques [186]. As noted before [548], there is a need for a comprehensive catalogue of feasibility counts for disorders and drugs in observational data sources. Such a catalogue will allow research teams to quickly estimate if a question is addressable and if

additional data sources are needed. Although there is an ongoing work on this topic [166,242], a comprehensive knowledge base does not exist yet.

Even if the data elements required by phenotyping algorithms are present, the study team has to inspect the accuracy of patient capture. While common phenotype validation methods such as manual chart review allow computing performance metrics to assess the quality of phenotyping algorithms [325], their use is not feasible due to time constraints. Instead, we assess the general plausibility of a cohort (e.g., number of patients, sequences of clinical events, sampling review of patient data) using data-driven approach. This approach is analogous to chart review in inspecting the features of groups of patients to determine if they belong to the studied populations and requires both data knowledge and clinical expertise. Patient profiles described in Section 3.5 were developed after we had ran the pilot. Otherwise, they could be a valuable addition to these practices.

Addressing bias

Appropriate methods to control for bias must be applied. When delivering evidence that intended to be used for clinical practice, the study team must assess evidence validity and reliability. Using methods that were previously shown to mitigate bias (large-scale propensity score, negative and positive controls) [204,530,537], as well as controlling for patient selection, detection, measurement and reporting bias, ensures that only high-quality evidence is delivered. First, study design and result interpretation require collaborative work of clinicians, statisticians and informaticians. Using a large-scale propensity score model as adopted by OHDSI [530] based on all demographic, condition, drug, measurement and procedure codes available in the structured data allows to achieve better control of confounding [208]. Using a large number of negative and synthetic positive controls allows to estimate the extent of residual bias present after statistical adjustment, and empirically calibrate the findings to

account for this systematic error [204]. Replicating the analyses on multiple databases allows to assess consistency of the results. By following a consistent approach to study design and execution and using standardized open-source analytic tools throughout the process, we also increase the transparency of study results, which is particularly important when study results raise concerns. For example, we observed that patients with end stage renal disorder on warfarin were substantially clinically different from patients on oral anticoagulants, so we could not achieve full propensity score balance needed to produce reliable estimates. Delivering such results require extensive and transparent communication with the target users regarding limitations of the study and potential bias.

As we used data transformed into a common data model, data processing and standardization have to be conducted beforehand by a separate extract-transform-load (ETL) team. While it means that additional data elements in the native EHR system are rarely used, the CDM allows us to minimize time spent on obtaining relevant information and address questions from various specialties. Moreover, the OHDSI data network provides an opportunity to leverage multiple data sources, increase sample size and include diverse populations.

Having data with different provenance also provides opportunities to select a data source appropriate for a specific research question. As previously shown, EHR data sources provide better capture of inpatient drug administration [88]. While we also observed this pattern, we noted that outpatient prescriptions were better captured in administrative claims datasets.

Nevertheless, EHR sources provided an opportunity to use laboratory test results and vital signs not otherwise available in administrative claims datasets. The former was crucial when studying underreported conditions as it allowed detecting patients of interest by using alternative laboratory criteria. Additionally, data source use may be prioritized based on predominant populations. For

example, in the question related to diabetes therapy (Appendix 5.6, question 8), we used MDCR as it both provided better capture of drug prescription frequency and duration and mainly had elderly patients with increased prevalence of diabetes mellitus type II. Nevertheless, while we can use multiple data sources converted to OMOP CDM in our institution, new policies and practices are needed to enable timely and seamless data exchange across institutions.

Finally, the last challenge was related to clinicians' perception of results. In concordance with the literature [549] , clinicians are likely to use informal reasoning in their decision-making. We observed that clinicians in our study were more inclined to use and disseminate our reports if the latter aligned with their baseline expectations.

We show that observational data can be used to generate evidence at the bedside. Nevertheless, future work needs to address several considerations to increase the usability of such services. First, increasing patient sample size by including other OMOP data sources into analysis would facilitate large scale propensity score comparison and enable research for rare outcomes. Second, using alternative approaches for phenotype performance estimation such as probabilistic evaluation [115] or patient profile review as described in Chapter 2 would enable better patient phenotyping. Third, as clinical rounds leveraged the most questions per clinician and allowed answering patient-relevant questions in a timely manner, future development can attempt to optimize clinician engagement and make it more scalable.

5.4 Chapter summary and lessons learned

In Chapters 3 and 4 we investigated pre-analysis bias and measurement error associated with patient phenotyping and comparator definition. We developed standardized pipelines, new methods and approaches for reducing bias, which, coupled with common environment (a common data model and standardized vocabularies) and robust large-scale analytical and statistical methods (large-scale propensity modelling, negative controls and p-value and confidence interval calibration), facilitate generation of reliable and robust evidence. While it is an achievement on its own, these standardized approaches also make the generation process more efficient and less time-consuming.

This opens a possibility of generating evidence in real or near-real time to address clinicians' immediate information needs.

As we observed throughout the interviews with clinicians, there are multiple, oftentimes shared, questions regarding drug effectiveness, safety and utilization that are not covered by the existing guidelines, clinical trials, or observational studies. This gap in clinical knowledge in not likely to go away with a growing number of new drugs highlighting a need for responsive yet reliable evidence.

Yet, it is not an easy endeavor. In our review of the tools attempting a similar task, only one attempted to properly address bias. Our experience with the Data Consult Service mirrored this finding at the main bottleneck in evidence generation and delivery was bias mitigation. To deliver reliable on-demand evidence, we had to ensure that the study design reflects clinicians' rationale, the concept sets, and phenotypes identify the patients of interest, the data quality issues are accounted for, the study diagnostics meet objective criteria of evidence quality. If our procedures showed potential bias at any of these steps, we would iterate over the previous steps. Not surprisingly, while most of the questions could be addressed within two days, some of them took weeks.

Having both clinical and informatics expertise was crucial in fast evidence delivery. Green Button, which was the predecessor of the Data Consult Service, had a team of data scientists, clinicians and local experts. We found that these aspects of expertise can effectively be combined in one person (the author of this thesis), which provides more flexibility and increases efficiency. As we noted, it is only possible if we rely on already established practices and tools for reliable evidence generation.

Although one may find these results discouraging, the efforts to produce reliable evidence resulted in high appreciation of the results by the clinicians and increased their trust in observational data.

While more research is needed to establish fast and efficient pipelines for evidence generation, the experience of the Data Consult Service encourages us in future efforts. We do not claim that observational data at a given institution can answer all clinical questions reliably, but as we accumulate more knowledge about the data and its limitations, we can both discard the questions we cannot answer faster and further develop targeted methods to address these limitations.

Chapter 6. Conclusions and future work

6.1. Conclusions

In this thesis, we hypothesized that actionable observational evidence can be generated in near-real time to address clinicians' information needs not covered by the existing evidence.

For such evidence to be used in decision-making for patient care, it must be reliable and accurate. To cover the gaps in methods for reliable observational evidence generation, we systematically assessed pre-analysis bias in observational studies and developed novel approaches and informatics tools to mitigate it.

Phenotyping or identification of patients of interest is a major source of measurement error in observational studies. As it can only be partially controlled by analytical methods, investigators are required to use all appropriate methods to ensure that patient capture is accurate. While phenotyping has been extensively studied over the years, the problem of developing scalable phenotypes that can be generalizable to multiple data sources has not been solved. To our knowledge, we were the first researchers to perform large-scale assessment of heterogeneity in observational data networks and its implications on phenotyping. We developed novel methods to estimate data source granularity, which, along with the analysis of real-world code utilization patterns, contributed to characterization of 22 US and international data sources containing more than 272 billion records.

This work informed development and evaluation of the recommender system (PHOEBE) that enables creating comprehensive concept sets that account for data source heterogeneity. As opposed to the concept sets developed on a local data source or those borrowed from the literature, concept sets generated with PHOEBE are generalizable to multiple data sources and can identify patients accurately

and early on in the course of the disease. PHOEBE substantially reduces the time needed to create portable concept sets even if the researchers do not have the access to all of them.

Similarly to the choice of codes, the choice of inclusion and exclusion criteria and the order of their application influences patient selection and may introduce bias. In our systematic experiment of assessing algorithmic implementations of the conceptual definition of a cohort, we highlighted significant variability that impacts patient composition. Explicit documentation and reporting in reproducible form of machine-readable objects is a necessary step in any observational study and is required for accurate interpretation of its results.

As phenotypes can introduce measurement error they should be evaluated in a robust fashion. The current gold standard for phenotype evaluation (manual chart review) is not always possible, is variable and is time-consuming. We developed and evaluated a novel data-driven and interpretable approach that organizes structured data in a systematic fashion using clinical reasoning principles. We demonstrated that our system (KEEPER) achieves similar accuracy of evaluation, better inter-rater reliability, and substantially faster review compared to manual chart review. KEEPER can enable scalable and reliable phenotyping on EHR and claims data sources.

Comparator definition is another source of bias in observational studies, which, as we demonstrated in this thesis, influences both patient composition and study estimates. We discovered lack of empirical evidence and guidance on how unexposed comparators should be defined. Variable unexposed comparator definition strategies can bias inference in safety and effectiveness studies or delay them, which is especially important for mass and urgent campaigns such as COVID-19 vaccination.

To cover this gap, we systematically assessed robustness of background incidence rates commonly used in safety surveillance. To our knowledge, we were the first to do a systematic and comprehensive examination of the magnitude of variability of incidence rates depending on different design choices. Observed high age- and gender-dependent variability in rates was expected, but sensitivity of background rates to the choice of data source, time-at-risk parameters, and unexposed comparator index date was not quantified at scale before. As those parameters highly influence the rates of events in general population and, as a result, interpretation of drug safety, regulatory bodies and researchers should carefully consider these choices when interpreting observed-to-expected studies.

We uncovered a large effect of index date selection strategy (anchoring) on the rates of events and baseline patient characteristics. We developed empirical methods to define an appropriate anchor to reflect target exposure features. If these or similar methods are not used, imperfect anchoring strategy can lead to biased estimates in cohort studies and self-controlled case series, especially if statistical methods fail to incorporate a large number of potential confounders to balance the target and comparator.

We used the background incidence rate methods that we developed to calculate rates for the European Medicines Agency during its determination on the AstraZeneca COVID-19 vaccine in March 2021. The Agency used our rates to reveal an increase in very rare clotting disorders, but no increase in more common clotting disorders, and the Agency determined that the vaccine's benefits outweighed its risks, making it available again throughout Europe.

We examined bias in vaccine effectiveness studies. As there was no consensus on interpretation of high COVID-19 vaccine effectiveness during the first two weeks after vaccination, we investigated short-term effectiveness and bias associated with it using robust analytical methods, a set of secondary

analyses and chart review. We found selection and health-seeking behavior biases as well as confounding by severity and indication that explained high effectiveness and provided further considerations for future short-term effectiveness studies.

We subsequently tested the hypothesis that building scalable and robust pipelines to address bias enables both more robust and faster evidence generation. Using the methods and knowledge we generated in Chapter 3 and 4 we proceeded with building a service that delivers actionable, reliable, and timely evidence to clinicians. We first explored the potential clinicians' questions newly generated evidence can address. We interviewed 31 clinicians at CUIMC and created a modern taxonomy of information needs not covered by the existing evidence. Such needs did not decrease with years of clinical experience and were mainly related to selecting an optimal treatment for patients not typically covered by RCTs, which highlighted a need for timely observational evidence.

We thoroughly analyzed the experience of previous CDSS that generated new evidence. As we characterized and classified 25 tools we found, we discovered that all available tools lacked robust and thorough bias-mitigating strategies. Based on the assessment of clinicians' information needs and review of existing clinical decision support tools, we designed a service (Data Consult Service) for generating new evidence at the point of care, which specifically focuses on providing robust and reliable evidence. We ran a pilot study at NewYork-Presbyterian hospital, prospectively collected 29 questions from 22 clinicians and answered 24 of them. On average, we were able to produce publishable quality reports in 1-2 days. We classified the challenges we encountered and proposed future steps for reliable real time evidence generation.

Summary

In this thesis, we showed that it is feasible to deliver newly generated on-demand observational evidence to clinicians. As such evidence is used in decision-making, it is mandatory to ensure its reliability and report its limitations transparently and honestly. These two principles increase trust in the evidence, which is especially important if clinicians are to act on it. Reliability requires systematic and standardized approaches to data quality assurance, patient phenotyping, comparator definition, and robust effect estimation.

Upon examination of the current state of the field of observational research, we identified the gaps in knowledge and lack of informatics solutions for patient phenotyping and comparator definition in heterogeneous distributed data networks. Throughout this thesis, we contributed, both methodologically and empirically, to addressing phenotyping and comparator definition as a source of pre-analysis bias, which resulted in more scalable and robust evidence generation pipelines.

The lessons we learned about data source heterogeneity and granularity, the impact of data standardization and harmonization, the ability of structured data to effectively reflect patient state, and the importance of addressing temporal, selection and other biases at the design stage were crucial to delivering evidence to the clinicians at NYP. The lessons we learned about the limitations of observational data may be even more important as they allowed us to discard unreliable evidence.

The evidence we generated throughout this thesis had direct clinical impact: we calculated background rates for adverse events of special interest influencing safety decision for 700 million patients and we generated the evidence that directly impacted patient care for 22 clinicians. We believe that our experience with the Data Consult Service can be scaled to a larger group of clinicians and a

larger group of institutions if we continue working on establishing scalable and standardized workflows for evidence generation.

6.3. Future directions

The findings reported in this thesis point to many directions for future research. Here, we present some of the possible extensions of this work in the order they appear in the thesis.

First, while we developed methods and tools to improve scalability, more can be done to establish more scalable and robust pipelines for phenotyping.

The methods for estimating data source granularity we developed can potentially inform the initial stages of phenotyping. Our methods used SNOMED-CT terms from the Condition domain, which represent a large portion of all structured data in a data source. Future studies may examine the hypothesis whether codes for procedures and measurements contain uncaptured information about granularity. Examining influence of procedural and measurement terminologies such as CPT-4 or LOINC on data source granularity would require aligning these terminologies to create joint hierarchies of procedures and measurements as the latter is currently lacking comprehensiveness in the OHDSI Standardized Vocabularies and does not exist elsewhere. We may also want to examine if leveraging other SNOMED-CT relationships that define concepts (such as ‘part-of’ or ‘has-a’) achieves better precision in estimating individual concept granularity to be used in the overall data source granularity estimation.

Another part of future work concerns concept selection in phenotyping. While PHOEBE, a recommender system for concept selection we developed, is efficient in expanding concept sets and is actively used in the network studies in OHDSI, more can be done to improve its recommendations. Adding data-driven approaches that leverage patient context can potentially enrich recommendations

[331,550]. As data-driven recommendations may produce more sensitive (broad) concept sets at a cost of decreased specificity, more studies are needed to determine an appropriate balance between more and less specific recommendations for efficient concept selection.

We will be doing more work on phenotype evaluation. We envision that our profile review system (KEEPER) can greatly contribute to scalable phenotype pipelines if integrated into a larger stack of OHDSI tools. To achieve that, we need to (a) build an executable package or user interface and (b) develop or adopt automated disease-agnostic approaches to relevant information extraction. There are many works on identifying similar concepts, including lexical, ontological and data-driven approaches [329–331] to can be leveraged to accomplish this task. An approach suitable for KEEPER will have to identify relevant but not necessarily semantically similar concepts, concepts from different domains (such as laboratory tests relevant to a given disease) and clinically meaningful concept pairs (such as diagnosis-differential diagnosis pairs [334]) and is therefore likely to be a mixed-methods approach.

More studies are needed to examine portability of patient profiles to institutions with different data capture or patient composition.

Second, more can be done to standardize comparator definition process and assess bias at this stage. We provided our recommendations on how to empirically select comparators and the next step can be formalizing the process and criteria so that they can become an executable package to guide comparator definition. Future work may concern investigating and establishing computational criteria to automate comparator selection. While we focused on the unexposed comparator, we hope to collaborate with our colleagues to expand this work to the exposed comparator definition and selection.

We hope that the future work (our and our colleagues) would enable us to scale the Data Consult Service to cover more clinicians, further reduce time and increase the scope of the questions we can answer.

For example, we can investigate if increasing patient sample size by including other OMOP data sources into analysis facilitates question answering. While it may enable research for rare outcomes, it requires establishing fast pipelines for study execution between institutions. Similarly, adding the currently missing elements to our OMOP CDM (such as blood culture results) may enable more efficient and faster evidence generation for those questions that require these elements.

Chapter 7. Bibliography

1 Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods Inf Med* 1998;**37**:394–403. doi:10.1055/s-0038-1634558

2 Mo H, Thompson WK, Rasmussen LV, *et al.* Desiderata for computable representations of electronic health records-driven phenotype algorithms. *Journal of the American Medical Informatics Association* 2015;**22**:1220–30. doi:10.1093/jamia/ocv112

3 Interoperability in Healthcare | HIMSS. https://www.himss.org/resources/interoperability-healthcare (accessed 17 Sep 2022).

4 Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *Journal of the American Medical Informatics Association* 2018;**25**:289–94. doi:10.1093/jamia/ocx110

5 Richesson RL, Sun J, Pathak J, *et al.* Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial Intelligence in Medicine* 2016;**71**:57–61. doi:10.1016/j.artmed.2016.05.005

6 Norman D. *Things that make us smart.* Basic Books 1994.

7 Ostropolets A, Reich C, Ryan P, *et al.* Characterizing database granularity using SNOMED-CT hierarchy. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association 2020. 983.

8 Ostropolets A, Ryan P, Hripcsak G. Phenotyping in distributed data networks: selecting the right codes for the right patients. In: *AMIA Annual Symposium Proceedings*. 2022.

9 Ostropolets A, Reich C, Ryan P, *et al.* Adapting electronic health records-derived phenotypes to claims data: Lessons learned in using limited clinical data for phenotyping. *Journal of Biomedical Informatics* 2020;**102**:103363. doi:10.1016/j.jbi.2019.103363

10 Ostropolets A, Li X, Makadia R, *et al.* Factors Influencing Background Incidence Rate Calculation: Systematic Empirical Evaluation Across an International Network of Observational Databases. *Front Pharmacol* 2022;**13**:814198. doi:10.3389/fphar.2022.814198

11 Ostropolets A, Ryan PB, Schuemie MJ, *et al.* Characterizing Anchoring Bias in Vaccine Comparator Selection Due to Health Care Utilization With COVID-19 and Influenza: Observational Cohort Study. *JMIR Public Health Surveill* 2022;**8**:e33099. doi:10.2196/33099

12 Ostropolets A, Hripcsak G. COVID-19 vaccination effectiveness rates by week and sources of bias: a retrospective cohort study. *BMJ Open* 2022;**12**:e061126. doi:10.1136/bmjopen-2022-061126

13 Ostropolets A, Chen R, Zhang L, *et al.* Characterizing physicians' information needs related to a gap in knowledge unmet by current evidence. *JAMIA Open* 2020;**3**:281–9. doi:10.1093/jamiaopen/ooaa012

14 Ostropolets A, Zhang L, Hripcsak G. A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. *Journal of the American Medical Informatics Association* Published Online First: 29 October 2020. doi:10.1093/jamia/ocaa200

15 Ostropolets A, Zachariah P, Ryan P, *et al.* Data Consult Service: Can we use observational data to address immediate clinical needs? *Journal of the American Medical Informatics Association* 2021;**28**:2139–46. doi:10.1093/jamia/ocab122

16 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;:n1435. doi:10.1136/bmj.n1435

17 Ostropolets A, Elias PA, Reyes MV, *et al.* Metformin Is Associated With a Lower Risk of Atrial Fibrillation and Ventricular Arrhythmias Compared With Sulfonylureas: An Observational Study. *Circ: Arrhythmia and Electrophysiology* 2021;**14**. doi:10.1161/CIRCEP.120.009115

18 Burn E, You SC, Sena AG, *et al.* Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nature communications* 2020;**11**:1–11.

19 Castano VG, Spotnitz M, Waldman GJ, *et al.* Identification of patients with drug resistant epilepsy in electronic medical record data using the Observational Medical Outcomes Partnership Common Data Model. *Epilepsia* 2022;:epi.17409. doi:10.1111/epi.17409

20 Chen R, Schuemie MJ, Suchard MA, *et al.* Evaluation of Large-scale Propensity Score Modeling and Covariate Balance on Potential Unmeasured Confounding in Observational Research. In: *AMIA Annual Symposium Proceedings*. 2020.

21 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. *BMJ Open* 2022;**12**:e057977. doi:10.1136/bmjopen-2021-057977

22 Kostka K, Duarte-Salles T, Prats-Uribe A, *et al.* Unraveling COVID-19: A Large-Scale Characterization of 4.5 Million COVID-19 Cases Using CHARYBDIS. *CLEP* 2022;**Volume 14**:369–84. doi:10.2147/CLEP.S323292

23 Lane JC, Weaver J, Kostka K, *et al.* Risk of depression, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multinational network cohort study. *Rheumatology* 2020.

24 Lane JCE, Weaver J, Kostka K, *et al.* Safety of hydroxychloroquine, alone and in combination with azithromycin, in light of rapid wide-spread use for COVID-19: a multinational, network cohort and self-controlled case series study. Rheumatology 2020. doi:10.1101/2020.04.08.20054551

25 Li X, Lai LY, Ostropolets A, *et al.* Bias, precision and timeliness of historical (background) rate comparison methods for vaccine safety monitoring: an empirical multi-database analysis. *Frontiers in Pharmacology* 2021;:3307.

26 Morales DR, Ostropolets A, Lai L, *et al.* Characteristics and outcomes of COVID-19 patients with and without asthma from the United States, South Korea, and Europe. *Journal of Asthma* 2022;:1–11. doi:10.1080/02770903.2021.2025392

27 Moreno-Martos D, Verhamme K, Ostropolets A, *et al.* Characteristics and outcomes of COVID-19 patients with COPD from the United States, South Korea, and Europe. *Wellcome Open Res* 2022;**7**:22. doi:10.12688/wellcomeopenres.17403.2

28 Reps JM, Kim C, Williams RD, *et al.* Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study. *JMIR medical informatics* 2021;**9**:e21547–e21547.

29 Reyes C, Pistillo A, Fernández-Bertolín S, *et al.* Characteristics and outcomes of patients with COVID-19 with and without prevalent hypertension: a multinational cohort study. *BMJ Open* 2021;**11**:e057632. doi:10.1136/bmjopen-2021-057632

30 Rodriguez V. Phenotype Concept Set Construction from Concept Pair Likelihoods. In: *AMIA 2020 Proceedings*. 2020.

31 Roel E, Pistillo A, Recalde M, *et al.* 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* 2021;**30**:1884–94. doi:10.1158/1055-9965.EPI-21-0266

32 Seong Y, You SC, Ostropolets A, *et al.* Incorporation of Korean Electronic Data Interchange Vocabulary into Observational Medical Outcomes Partnership Vocabulary. *Healthcare Informatics Research* 2021;**27**:29–38.

33 Shoaibi A, Rao GA, Voss EA, *et al.* Phenotype Algorithms for the Identification and Characterization of Vaccine-Induced Thrombotic Thrombocytopenia in Real World Data: A Multinational Network Cohort Study. *Drug Saf* 2022;**45**:685–98. doi:10.1007/s40264-022-01187-y

34 Spotnitz M, Ostropolets A, Castano VG, *et al.* Patient characteristics and antiseizure medication pathways in newly diagnosed epilepsy: Feasibility and pilot results using the common data model in a single-center electronic medical record database. *Epilepsy & Behavior* 2022;**129**:108630. doi:10.1016/j.yebeh.2022.108630

35 Tan EH, Sena AG, Prats-Uribe A, *et al.* COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries. *Rheumatology* 2021;**60**:SI37–50. doi:10.1093/rheumatology/keab250

36 Sackett DL, Rosenberg WMC, Gray JAM, *et al.* Evidence based medicine: what it is and what it isn't. *BMJ* 1996;**312**:71–2. doi:10.1136/bmj.312.7023.71

37 Timmermans S, Mauck A. The Promises And Pitfalls Of Evidence-Based Medicine. *Health Affairs* 2005;**24**:18–28. doi:10.1377/hlthaff.24.1.18

38 Chow N, Gallo L, Busse JW. Evidence-based medicine and precision medicine: Complementary approaches to clinical decision-making. *Precision Clinical Medicine* 2018;**1**:60–4. doi:10.1093/pcmedi/pby009

39 Bahtsevani C, Udén G, Willman A. Outcomes of evidence-based clinical practice guidelines: A systematic review. *International Journal of Technology Assessment in Health Care* 2004;**20**:427–33. doi:10.1017/S026646230400131X

40 Cook DA, Sorensen KJ, Wilkinson JM, *et al.* Barriers and decisions when answering clinical questions at the point of care: a grounded theory study. *JAMA Intern Med* 2013;**173**:1962–9. doi:10.1001/jamainternmed.2013.10103

41 Brown PJ, Borowitz SM, Novicoff W. Information exchange in the NICU: what sources of patient data do physicians prefer to use? *International Journal of Medical Informatics* 2004;**73**:349–55. doi:10.1016/j.ijmedinf.2004.03.001

42 Glasgow RE, Green LW, Klesges LM, *et al.* External validity: we need to do more. *Annals of Behavioral Medicine* 2006;**31**:105–8.

43 Kennedy-Martin T, Curtis S, Faries D, *et al.* A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015;**16**. doi:10.1186/s13063-015-1023-4

44 Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science* 2015;**16**:475–85. doi:10.1007/s11121-014-0513-z

45 Boyd CM, Vollenweider D, Puhan MA. Informing Evidence-Based Decision-Making for Patients with Comorbidity: Availability of Necessary Information in Clinical Trials for Chronic Diseases. *PLoS ONE* 2012;**7**:e41601. doi:10.1371/journal.pone.0041601

46 Schmidt AF, Groenwold RHH, van Delden JJM, *et al.* Justification of exclusion criteria was underreported in a review of cardiovascular trials. *Journal of Clinical Epidemiology* 2014;**67**:635–44. doi:10.1016/j.jclinepi.2013.12.005

47 Bell SA, Tudur Smith C. A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of ClinicalTrials.gov. *Orphanet Journal of Rare Diseases* 2014;**9**. doi:10.1186/s13023-014-0170-0

48 Griggs RC, Batshaw M, Dunkle M, *et al.* Clinical research for rare disease: Opportunities, challenges, and solutions. *Molecular Genetics and Metabolism* 2009;**96**:20–6. doi:10.1016/j.ymgme.2008.10.003

49 Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern Med* 2014;**174**:710–8. doi:10.1001/jamainternmed.2014.368

50 Daei A, Soleymani MR, Ashrafi□rizi H, *et al.* Personal, technical and organisational factors affect whether physicians seek answers to clinical questions during patient care: a literature review. *Health Information & Libraries Journal* Published Online First: 20 July 2020. doi:10.1111/hir.12323

51 Ely JW, Osheroff JA, Ebell MH, *et al.* Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999;**319**:358–61. doi:10.1136/bmj.319.7206.358

52 Smith R. What clinical information do doctors need? *BMJ* 1996;**313**:1062–8. doi:10.1136/bmj.313.7064.1062

53 Karnon J, Partington A, Horsfall M, *et al.* Variation in Clinical Practice: A Priority Setting Approach to the Staged Funding of Quality Improvement. *Applied Health Economics and Health Policy* 2016;**14**:21–7. doi:10.1007/s40258-015-0160-y

54 Siemieniuk RA, Bartoszko JJ, Ge L, *et al.* Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ* 2020;:m2980. doi:10.1136/bmj.m2980

55 Neupane NP, Das AK, Singh AK, *et al.* Off Label Medication to Combat COVID-19: Review Results to Date. *COVID* 2021;**2**:496–506. doi:10.2174/2666796701999200729145342

56 Visvanathan K, Levit LA, Raghavan D, *et al.* Untapped Potential of Observational Research to Inform Clinical Decision Making: American Society of Clinical Oncology Research Statement. *J Clin Oncol* 2017;**35**:1845–54. doi:10.1200/JCO.2017.72.6414

57 Research C for DE and. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. U.S. Food and Drug Administration. 2021.https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory (accessed 18 Oct 2022).

58 Wang SV, Sreedhara SK, Schneeweiss S, *et al.* Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat Commun* 2022;**13**:5126. doi:10.1038/s41467-022-32310-3

59 Nosek BA, Hardwicke TE, Moshontz H, *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. *Annu Rev Psychol* 2022;**73**:719–48. doi:10.1146/annurev-psych-020821-114157

60 Grimberg F, Asprion PM, Schneider B, *et al.* The Real-World Data Challenges Radar: A Review on the Challenges and Risks regarding the Use of Real-World Data. *Digit Biomark* 2021;**5**:148–57. doi:10.1159/000516178

61 Burns L, Roux NL, Kalesnik-Orszulak R, *et al.* Real-World Evidence for Regulatory Decision-Making: Guidance From Around the World. *Clinical Therapeutics* 2022;**44**:420–37. doi:10.1016/j.clinthera.2022.01.012

62 Liu M, Qi Y, Wang W, *et al.* Toward a better understanding about real-world evidence. *Eur J Hosp Pharm* 2022;**29**:8–11. doi:10.1136/ejhpharm-2021-003081

63 Gokhale M, Stürmer T, Buse JB. Real-world evidence: the devil is in the detail. *Diabetologia* 2020;**63**:1694–705. doi:10.1007/s00125-020-05217-1

64 Rothman KJ, Greenland S, editors. *Modern epidemiology*. 2nd ed. Philadelphia, PA: : Lippincott-Raven 1998.

65 Suissa S, Dell'Aniello S. Time□related biases in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf* 2020;**29**:1101–10. doi:10.1002/pds.5083

66 Newcomer SR, Kulldorff M, Xu S, *et al.* Bias from outcome misclassification in immunization schedule safety research. *Pharmacoepidemiol Drug Saf* 2018;**27**:221–8. doi:10.1002/pds.4374

67 Grimes DA, Schulz KF. Bias and causal associations in observational research. *The Lancet* 2002;**359**:248–52. doi:10.1016/S0140-6736(02)07451-2

68 Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Springer 2009.

69 De Smedt T, Merrall E, Macina D, *et al.* Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. *PLOS ONE* 2018;**13**:e0199180. doi:10.1371/journal.pone.0199180

70 Gagne JJ, Fireman B, Ryan PB, *et al.* Design considerations in an active medical product safety monitoring system: DESIGN CONSIDERATIONS FOR ACTIVE MONITORING. *Pharmacoepidemiol Drug Saf* 2012;**21**:32–40. doi:10.1002/pds.2316

71 Remschmidt C, Wichmann O, Harder T. Frequency and impact of confounding by indication and healthy vaccinee bias in observational studies assessing influenza vaccine effectiveness: a systematic review. *BMC Infect Dis* 2015;**15**:429. doi:10.1186/s12879-015-1154-y

72 Shang N, Liu C, Rasmussen LV, *et al.* Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. *Journal of Biomedical Informatics* 2019;**99**:103293. doi:10.1016/j.jbi.2019.103293

73 Shinde MU, Baker MA, Spencer-Smith C, *et al.* Validation of transfusion administrations among potential Transfusion-Related Acute Lung Injury (TRALI) patients included in the Sentinel Distributed Database. ;:43.

74 Schuemie MJ, Ryan PB, Pratt N, *et al.* Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *Journal of the American Medical Informatics Association* 2020;**27**:1331–7. doi:10.1093/jamia/ocaa103

75 Berger ML, Sox H, Willke RJ, *et al.* Good Practices for Real□World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR□ISPE Special Task Force on Real□World Evidence in Health Care Decision Making. *Value in Health* 2017;**20**:1003–8. doi:10.1016/j.jval.2017.08.3019

76 Knottnerus A, Tugwell P. STROBE—A checklist to STrengthen the Reporting of OBservational studies in Epidemiology. *Journal of Clinical Epidemiology* 2008;**61**:323. doi:10.1016/j.jclinepi.2007.11.006

77 Wang SV, Schneeweiss S, Berger ML, *et al.* Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. *Value in Health* 2017;**20**:1009–22. doi:10.1016/j.jval.2017.08.3018

78 Burns PB, Rohrich RJ, Chung KC. The Levels of Evidence and Their Role in Evidence-Based Medicine: *Plastic and Reconstructive Surgery* 2011;**128**:305–10. doi:10.1097/PRS.0b013e318219c171

79 Van Spall HGC, Toren A, Kiss A, *et al.* Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review. *JAMA* 2007;**297**:1233. doi:10.1001/jama.297.11.1233

80 Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet (London, England)* 2005;**365**:82–93. doi:10.1016/S0140-6736(04)17670-8

81 Marchal B, Westhorp G, Wong G, *et al.* Realist RCTs of complex interventions – An oxymoron. *Social Science & Medicine* 2013;**94**:124–8. doi:10.1016/j.socscimed.2013.06.025

82 Ely JW, Osheroff JA, Ebell MH, *et al.* Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999;**319**:358–61. doi:10.1136/bmj.319.7206.358

83 Dreyer NA, Tunis SR, Berger M, *et al.* Why Observational Studies Should Be Among The Tools Used In Comparative Effectiveness Research. *Health Affairs* 2010;**29**:1818–25. doi:10.1377/hlthaff.2010.0666

84 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 2013;**20**:117–21. doi:10.1136/amiajnl-2012-001145

85 Giordano C, Brennan M, Mohamed B, *et al.* Accessing Artificial Intelligence for Clinical Decision-Making. *Front Digit Health* 2021;**3**:645232. doi:10.3389/fdgth.2021.645232

86 Frieden TR. Evidence for Health Decision Making — Beyond Randomized, Controlled Trials. *New England Journal of Medicine* 2017;**377**:465–75. doi:10.1056/NEJMra1614394

87 Eichler H, Baird L, Barker R, *et al.* From adaptive licensing to adaptive pathways: Delivering a flexible life□span approach to bring new drugs to patients. *Clin Pharmacol Ther* 2015;**97**:234–46. doi:10.1002/cpt.59

88 Lin K, Schneeweiss S. Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. *Clinical Pharmacology & Therapeutics* 2016;**100**:147–59. doi:10.1002/cpt.359

89 Riva JJ, Malik KMP, Burnie SJ, *et al.* What is your research question? An introduction to the PICOT format for clinicians. *J Can Chiropr Assoc* 2012;**56**:167–71.

90 Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide | Effective Health Care (EHC) Program. https://effectivehealthcare.ahrq.gov/products/observational-cer-protocol/research (accessed 20 Oct 2022).

91 Lai LY, Arshad F, Areia C, *et al.* Current Approaches to Vaccine Safety Using Observational Data: A Rationale for the EUMAEUS (Evaluating Use of Methods for Adverse Events Under Surveillance-for Vaccines) Study Design. *Front Pharmacol* 2022;**13**:837632. doi:10.3389/fphar.2022.837632

92 Schuemie MJ, Ryan PB, Man KKC, *et al.* A plea to stop using the case□control design in retrospective database studies. *Statistics in Medicine* 2019;**38**:4199–208. doi:10.1002/sim.8215

93 Reps JM, Ryan PB, Rijnbeek PR, *et al.* Design matters in patient-level prediction: evaluation of a cohort vs. case-control design when developing predictive models in observational healthcare datasets. *J Big Data* 2021;**8**:108. doi:10.1186/s40537-021-00501-2

94 Wilson ED, Clairon Q, Henderson R, *et al.* Dealing with observational data in control. *Annual Reviews in Control* 2018;**46**:94–106. doi:10.1016/j.arcontrol.2018.05.001

95 Chan KS, Fowles JB, Weiner JP. Review: Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature. *Med Care Res Rev* 2010;**67**:503–27. doi:10.1177/1077558709359007

96 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 (Generating Evidence & Methods to improve patient outcomes)* 2016;**4**:18. doi:10.13063/2327-9214.1244

97 Wrenn JO, Stein DM, Bakken S, *et al.* Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association* 2010;**17**:49–53. doi:10.1197/jamia.M3390

98 Data Quality Review and Characterization Programs | Sentinel Initiative. https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model/data-quality-review-and-characterization-programs (accessed 10 Oct 2022).

99 Brown JS, Kahn M, Toh S. Data Quality Assessment for Comparative Effectiveness Research in Distributed Data Networks. *Medical Care* 2013;**51**:S22–9. doi:10.1097/MLR.0b013e31829b1e2c

100 Huser V, Kahn MG, Brown JS, *et al.* Methods for examining data quality in healthcare integrated data repositories. In: *Biocomputing 2018*. Kohala Coast, Hawaii, USA: : WORLD SCIENTIFIC 2018. 628–33. doi:10.1142/9789813235533_0059

101 Informatics OHDS and. *Chapter 15 Data Quality | The Book of OHDSI*. https://ohdsi.github.io/TheBookOfOhdsi/ (accessed 8 Oct 2020).

102 Kahn MG, Brown JS, Chun AT, *et al.* Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMs* 2015;**3**:7. doi:10.13063/2327-9214.1052

103 Pezoulas VC, Kourou KD, Kalatzis F, *et al.* Medical data quality assessment: On the development of an automated framework for medical data curation. *Computers in Biology and Medicine* 2019;**107**:270–83. doi:10.1016/j.compbiomed.2019.03.001

104 Juárez D, Schmidt EE, Stahl-Toyota S, *et al.* A Generic Method and Implementation to Evaluate and Improve Data Quality in Distributed Research Networks. *Methods Inf Med* 2019;**58**:086–93. doi:10.1055/s-0039-1693685

105 Hansen RA, Zeng P, Ryan P, *et al.* Exploration of heterogeneity in distributed research network drug safety analyses: Heterogeneity in Distributed Data Networks. *Res Syn Meth* 2014;**5**:352–70. doi:10.1002/jrsm.1121

106 van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *International Journal of Epidemiology* 2020;**49**:338–47. doi:10.1093/ije/dyz251

107 Doyle CM, Lix LM, Hemmelgarn BR, *et al.* Data variability across Canadian administrative health databases: Differences in content, coding, and completeness. *Pharmacoepidemiol Drug Saf* 2020;**29**:68–77. doi:10.1002/pds.4889

108 Ross MK, Zhu B, Natesan A, *et al.* Accuracy of Asthma Computable Phenotypes to Identify Pediatric Asthma Cases at an Academic Institution. In: *B52. PEDIATRIC ASTHMA AND ALLERGY*. American Thoracic Society 2020. A3710–A3710. doi:10.1164/ajrccm-conference.2020.201.1_MeetingAbstracts.A3710

109 Chen Y. Opportunities and Challenges in Data-Driven Healthcare Research. MEDICINE & PHARMACOLOGY 2018. doi:10.20944/preprints201806.0137.v1

110 Griffiths EA, Macaulay R, Vadlamudi NK, *et al.* The Role of Noncomparative Evidence in Health Technology Assessment Decisions. *Value in Health* 2017;**20**:1245–51. doi:10.1016/j.jval.2017.06.015

111 Rasmussen LV, Kiefer RC, Mo H, *et al.* A Modular Architecture for Electronic Health Record-Driven Phenotyping. *AMIA Jt Summits Transl Sci Proc* 2015;**2015**:147–51.

112 Meineke FA, Stäubert S, Löbe M, *et al.* Design and Concept of the SMITH Phenotyping Pipeline. *Stud Health Technol Inform* 2019;**267**:164–72. doi:10.3233/SHTI190821

113 Almowil ZA, Zhou S-M, Brophy S. Concept Libraries for Automatic Electronic Health Record Based Phenotyping: A Review. *IJPDS* 2021;**6**. doi:10.23889/ijpds.v6i1.1362

114 Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association* 2013;**20**:e206–11. doi:10.1136/amiajnl-2013-002428

115 Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. *Journal of Biomedical Informatics* 2019;**97**:103258. doi:10.1016/j.jbi.2019.103258

116 Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*. 0 ed. Chapman and Hall/CRC 2010. doi:10.1201/9781420066586

117 Camplain R, Kucharska-Newton A, Cuthbertson CC, *et al.* Misclassification of incident hospitalized and outpatient heart failure in administrative claims data: the Atherosclerosis Risk in

Communities (ARIC) study: MISCLASSIFICATION OF HEART FAILURE. *Pharmacoepidemiol Drug Saf* 2017;**26**:421–8. doi:10.1002/pds.4162

118 Griffiths RI, O'Malley CD, Herbert RJ, *et al.* Misclassification of incident conditions using claims data: impact of varying the period used to exclude pre-existing disease. *BMC Med Res Methodol* 2013;**13**:32. doi:10.1186/1471-2288-13-32

119 Dagliati A, Geifman N, Peek N, *et al.* Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artificial Intelligence in Medicine* 2020;**108**:101930. doi:10.1016/j.artmed.2020.101930

120 Freemantle N, Marston L, Walters K, *et al.* Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;**347**:f6409–f6409. doi:10.1136/bmj.f6409

121 Psaty BM, Koepsell TD, Lin D, *et al.* Assessment and Control for Confounding by Indication in Observational Studies. *Journal of the American Geriatrics Society* 1999;**47**:749–54. doi:10.1111/j.1532-5415.1999.tb01603.x

122 Shrank WH, Patrick AR, Alan Brookhart M. Healthy User and Related Biases in Observational Studies of Preventive Interventions: A Primer for Physicians. *J GEN INTERN MED* 2011;**26**:546–50. doi:10.1007/s11606-010-1609-1

123 Remschmidt C, Wichmann O, Harder T. Frequency and impact of confounding by indication and healthy vaccinee bias in observational studies assessing influenza vaccine effectiveness: a systematic review. *BMC Infect Dis* 2015;**15**:429. doi:10.1186/s12879-015-1154-y

124 Matthews KA, Kuller LH, Wing RR, *et al.* Prior to Use of Estrogen Replacement Therapy, Are Users Healthier than Nonusers? *American Journal of Epidemiology* 1996;**143**:971–8. doi:10.1093/oxfordjournals.aje.a008678

125 Presley CA, Chipman J, Min JY, *et al.* Evaluation of Frailty as an Unmeasured Confounder in Observational Studies of Antidiabetic Medications. *The Journals of Gerontology: Series A* 2019;**74**:1282–8. doi:10.1093/gerona/gly224

126 Kahn R, Schrag SJ, Verani JR, *et al.* Identifying and Alleviating Bias Due to Differential Depletion of Susceptible People in Postmarketing Evaluations of COVID-19 Vaccines. *American Journal of Epidemiology* 2022;**191**:800–11. doi:10.1093/aje/kwac015

127 Lipsitch M, Goldstein E, Ray GT, *et al.* Depletion-of-susceptibles bias in influenza vaccine waning studies: how to ensure robust results. *Epidemiol Infect* 2019;**147**:e306. doi:10.1017/S0950268819001961

128 Suissa S. Immortal Time Bias in Pharmacoepidemiology. *American Journal of Epidemiology* 2008;**167**:492–9. doi:10.1093/aje/kwm324

129 Levenson MS. Regulatory-grade clinical trial design using real-world data. *Clinical Trials* 2020;**17**:377–82. doi:10.1177/1740774520905576

130 D'Arcy M, Stürmer T, Lund JL. The importance and implications of comparator selection in pharmacoepidemiologic research. *Curr Epidemiol Rep* 2018;**5**:272–83. doi:10.1007/s40471-018-0155-y

131 Setoguchi S, Gerhard T. Comparator selection. In: *Developing a protocol for observational comparative effectiveness research: a user's guide*. Agency for Healthcare Research and Quality (US) 2013.

132 Carnahan RM, Gagne JJ, Hampp C, *et al.* Evaluation of the US Food and Drug Administration Sentinel Analysis Tools Using a Comparator with a Different Indication: Comparing the Rates of Gastrointestinal Bleeding in Warfarin and Statin Users. *Pharm Med* 2019;**33**:29–43. doi:10.1007/s40290-018-00265-w

133 Farrington CP, Nash J, Miller E. Case Series Analysis of Adverse Reactions to Vaccines: A Comparative Evaluation. *American Journal of Epidemiology* 1996;**143**:1165–73. doi:10.1093/oxfordjournals.aje.a008695

134 Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 1984;**79**:516–24. doi:10.1080/01621459.1984.10478078

135 Arbogast PG, Ray WA. Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. *American Journal of Epidemiology* 2011;**174**:613–20. doi:10.1093/aje/kwr143

136 Greenland S, Robins JM. Confounding and misclassification. *American Journal of Epidemiology* 1985;**122**:495–506. doi:10.1093/oxfordjournals.aje.a114131

137 Leon DA. Failed or misleading adjustment for confounding. *The Lancet* 1993;**342**:479–81. doi:10.1016/0140-6736(93)91599-H

138 Franklin JM, Platt R, Dreyer NA, *et al.* When Can Nonrandomized Studies Support Valid Inference Regarding Effectiveness or Safety of New Medical Treatments? *Clin Pharma and Therapeutics* 2022;**111**:108–15. doi:10.1002/cpt.2255

139 Schuemie MJ, Ryan PB, Hripcsak G, *et al.* Improving reproducibility by using high-throughput observational studies with empirical calibration. *Phil Trans R Soc A* 2018;**376**:20170356. doi:10.1098/rsta.2017.0356

140 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015;**12**:e1001885. doi:10.1371/journal.pmed.1001885

141 Kurz X, Perez-Gutthann S, the ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: Ten years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). *Pharmacoepidemiol Drug Saf* 2018;**27**:245–52. doi:10.1002/pds.4381

142 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;**13**:1. doi:10.1186/s12916-014-0241-z

143 Coiera E, Ammenwerth E, Georgiou A, *et al.* Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association* 2018;**25**:963–8. doi:10.1093/jamia/ocy028

144 Harris JK, Johnson KJ, Carothers BJ, *et al.* Use of reproducible research practices in public health: A survey of public health analysts. *PLoS ONE* 2018;**13**:e0202447. doi:10.1371/journal.pone.0202447

145 Hardwicke TE, Wallach JD, Kidwell MC, *et al.* An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *R Soc open sci* 2020;**7**:190806. doi:10.1098/rsos.190806

146 Popovic JR. Distributed data networks: a blueprint for Big Data sharing and healthcare analytics: Distributed data networks. *Annals of the New York Academy of Sciences* 2017;**1387**:105–11. doi:10.1111/nyas.13287

147 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.

148 Ball R, Robb M, Anderson S, *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

149 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. *Pharmacoepidem Drug Safe* 2011;**20**:1–11. doi:10.1002/pds.2053

150 Glanz JM, Marcuse EK, Weintraub E, *et al.* White Paper on studying the safety of the childhood immunization schedule in the Vaccine Safety Datalink. *Vaccine* 2016;**34**:A1–29. doi:10.1016/j.vaccine.2015.10.082

151 and The eMERGE Network, Gottesman O, Kuivaniemi H, *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Medicine* 2013;**15**:761–71. doi:10.1038/gim.2013.72

152 Forrest CB, Margolis P, Seid M, *et al.* PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. *Health Aff (Millwood)* 2014;**33**:1171–7. doi:10.1377/hlthaff.2014.0127

153 Afshar M, Press VG, Robison RG, *et al.* A computable phenotype for asthma case identification in adult and pediatric patients: External validation in the Chicago Area Patient-Outcomes Research

Network (CAPriCORN). *Journal of Asthma* 2018;**55**:1035–42. doi:10.1080/02770903.2017.1389952

154 Murphy SN, Weber G, Mendis M, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association* 2010;**17**:124–30. doi:10.1136/jamia.2009.000893

155 All of Us Research Program Investigators, Denny JC, Rutter JL, *et al.* The "All of Us" Research Program. *N Engl J Med* 2019;**381**:668–76. doi:10.1056/NEJMsr1809937

156 Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association* 2021;**28**:427–43. doi:10.1093/jamia/ocaa196

157 Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *npj Digit Med* 2020;**3**:109. doi:10.1038/s41746-020-00308-0

158 Richesson RL, Hammond WE, Nahm M, *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory: Table 1. *J Am Med Inform Assoc* 2013;**20**:e226–31. doi:10.1136/amiajnl-2013-001926

159 Southworth MR, Reichman ME, Unger EF. Dabigatran and Postmarketing Reports of Bleeding. *N Engl J Med* 2013;**368**:1272–4. doi:10.1056/NEJMp1302834

160 Schuemie MJ, Hripcsak G, Ryan PB, *et al.* Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci USA* 2018;**115**:2571–7. doi:10.1073/pnas.1708282114

161 Garza M, Del Fiol G, Tenenbaum J, *et al.* Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics* 2016;**64**:333–41. doi:10.1016/j.jbi.2016.10.016

162 Schneeweiss S, Brown JS, Bate A, *et al.* Choosing Among Common Data Models for Real World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products. *Clin Pharmacol Ther* 2020;**107**:827–33. doi:10.1002/cpt.1577

163 Hripcsak G, Shang N, Peissig PL, *et al.* Facilitating phenotype transfer using a common data model. *Journal of Biomedical Informatics* 2019;**96**:103253. doi:10.1016/j.jbi.2019.103253

164 Hripcsak G, Levine ME, Shang N, *et al.* Effect of vocabulary mapping for conditions on phenotype cohorts. *Journal of the American Medical Informatics Association* 2018;**25**:1618–25. doi:10.1093/jamia/ocy124

165 Burns EM, Rigby E, Mamidanna R, *et al.* Systematic review of discharge coding accuracy. *Journal of Public Health* 2012;**34**:138–48. doi:10.1093/pubmed/fdr054

166 Kirby JC, Speltz P, Rasmussen LV, *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016;**23**:1046–52. doi:10.1093/jamia/ocv202

167 Ahmad FS, Ricket IM, Hammill BG, *et al.* Computable Phenotype Implementation for a National, Multicenter Pragmatic Clinical Trial: Lessons Learned From ADAPTABLE. *Circ: Cardiovascular Quality and Outcomes* 2020;**13**. doi:10.1161/CIRCOUTCOMES.119.006292

168 Schneeweiss S, Brown JS, Bate A, *et al.* Choosing Among Common Data Models for Real World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products. *Clin Pharmacol Ther* 2020;**107**:827–33. doi:10.1002/cpt.1577

169 Morley KI, Wallace J, Denaxas SC, *et al.* Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. *PLoS ONE* 2014;**9**:e110900. doi:10.1371/journal.pone.0110900

170 Chapman M, Rasmussen LV, Pacheco JA, *et al.* Phenoflow: A Microservice Architecture for Portable Workflow-based Phenotype Definitions. *AMIA Jt Summits Transl Sci Proc* 2021;**2021**:142–51.

171 Denaxas S, Gonzalez-Izquierdo A, Direk K, *et al.* UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *Journal of the American Medical Informatics Association* 2019;**26**:1545–59. doi:10.1093/jamia/ocz105

172 Cho K. Introduction to the VA Phenomics Library (VAPheLib). 2020.

173 Weaver J, Potvien A, Swerdel J. Best practices for creating the standardized content of an entry in the OHDSI phenotype library. Washington, D.C.: 2019. 46.

174 Zhang L, Zhang Y, Cai T, *et al.* Automated grouping of medical codes via multiview banded spectral clustering. *Journal of Biomedical Informatics* 2019;**100**:103322. doi:10.1016/j.jbi.2019.103322

175 Zhao J, Zhang Y, Schlueter DJ, *et al.* Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study. *Journal of Biomedical Informatics* 2019;**98**:103270. doi:10.1016/j.jbi.2019.103270

176 Chen Y, Ghosh J, Bejan CA, *et al.* Building bridges across electronic health record systems through inferred phenotypic topics. *Journal of Biomedical Informatics* 2015;**55**:82–93. doi:10.1016/j.jbi.2015.03.011

177 Lee J, Liu C, Kim JH, *et al.* Comparative effectiveness of medical concept embedding for feature engineering in phenotyping. *JAMIA Open* 2021;**4**:ooab028. doi:10.1093/jamiaopen/ooab028

178 Yu S, Chakrabortty A, Liao KP, *et al.* Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association* 2017;**24**:e143–9. doi:10.1093/jamia/ocw135

179 Hong C, Rush E, Liu M, *et al.* Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *npj Digit Med* 2021;**4**:151. doi:10.1038/s41746-021-00519-z

180 Yu Y, Ruddy KJ, Wen A, *et al.* Integrating Electronic Health Record Data into the ADEpedia-on-OHDSI Platform for Improved Signal Detection: A Case Study of Immune-related Adverse Events. *AMIA Jt Summits Transl Sci Proc* 2020;**2020**:710–9.

181 Yu W, Zheng C, Xie F, *et al.* The use of natural language processing to identify vaccine□related anaphylaxis at five health care systems in the Vaccine Safety Datalink. *Pharmacoepidemiol Drug Saf* 2020;**29**:182–8. doi:10.1002/pds.4919

182 Savova GK, Fan J, Ye Z, *et al.* Discovering peripheral arterial disease cases from radiology notes using natural language processing. *AMIA Annu Symp Proc* 2010;**2010**:722–6.

183 Liu S, Wen A, Wang L, *et al.* An Open Natural Language Processing Development Framework for EHR-based Clinical Research: A case demonstration using the National COVID Cohort Collaborative (N3C). Published Online First: 2021. doi:10.48550/ARXIV.2110.10780

184 Wang Y, Wang L, Rastegar-Mojarad M, *et al.* Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics* 2018;**77**:34–49. doi:10.1016/j.jbi.2017.11.011

185 Wyner Z, Dublin S, Chambers C, *et al.* The FDA MyStudies app: a reusable platform for distributed clinical trials and real-world evidence studies. *JAMIA Open* 2021;**3**:500–5. doi:10.1093/jamiaopen/ooaa061

186 Sharma H, Mao C, Zhang Y, *et al.* Developing a portable natural language processing based phenotyping system. *BMC Medical Informatics and Decision Making* 2019;**19**. doi:10.1186/s12911-019-0786-z

187 Park J, You SC, Jeong E, *et al.* A Framework (SOCRATex) for Hierarchical Annotation of Unstructured Electronic Health Records and Integration Into a Standardized Medical Database: Development and Usability Study. *JMIR Med Inform* 2021;**9**:e23983. doi:10.2196/23983

188 Carrell DS, Gruber S, Floyd JS, *et al.* Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. In: *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY*. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA 2021. 16–7.

189 Banda JM, Halpern Y, Sontag D, *et al.* Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc* 2017;**2017**:48–57.

190 Kashyap M, Seneviratne M, Banda JM, *et al.* Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. *Journal of the American Medical Informatics Association* 2020;**27**:877–83. doi:10.1093/jamia/ocaa032

191 Pacheco JA, Rasmussen LV, Kiefer RC, *et al.* A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record

environments. *Journal of the American Medical Informatics Association* 2018;**25**:1540–6. doi:10.1093/jamia/ocy101

192 Tison GH, Chamberlain AM, Pletcher MJ, *et al.* Identifying heart failure using EMR-based algorithms. *International Journal of Medical Informatics* 2018;**120**:1–7. doi:10.1016/j.ijmedinf.2018.09.016

193 Nadkarni GN, Gottesman O, Linneman JG, *et al.* Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc* 2014;**2014**:907–16.

194 Sohn S, Wang Y, Wi C-I, *et al.* Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association* 2018;**25**:353–9. doi:10.1093/jamia/ocx138

195 Wang Q, Reps JM, Kostka KF, *et al.* Development and validation of a prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale in the OHDSI network. *PLOS ONE* 2020;**15**:e0226718. doi:10.1371/journal.pone.0226718

196 Liu C, Ta CN, Rogers JR, *et al.* Ensembles of natural language processing systems for portable phenotyping solutions. *Journal of Biomedical Informatics* 2019;**100**:103318. doi:10.1016/j.jbi.2019.103318

197 Oellrich A, Collier N, Smedley D, *et al.* Generation of Silver Standard Concept Annotations from Biomedical Texts with Special Relevance to Phenotypes. *PLoS ONE* 2015;**10**:e0116040. doi:10.1371/journal.pone.0116040

198 Gibson TB, Nguyen MD, Burrell T, *et al.* Electronic phenotyping of health outcomes of interest using a linked claims-electronic health record database: Findings from a machine learning pilot

project. *Journal of the American Medical Informatics Association* 2021;**28**:1507–17. doi:10.1093/jamia/ocab036

199 Geva A, Liu M, Panickan VA, *et al.* A high-throughput phenotyping algorithm is portable from adult to pediatric populations. *Journal of the American Medical Informatics Association* 2021;**28**:1265–9. doi:10.1093/jamia/ocaa343

200 Liu C, Zeinomar N, Chung WK, *et al.* Generalizability of Polygenic Risk Scores for Breast Cancer Among Women With European, African, and Latinx Ancestry. *JAMA Netw Open* 2021;**4**:e2119084. doi:10.1001/jamanetworkopen.2021.19084

201 Venkatesh KK, Castro VM, Perlis RH, *et al.* Impact of antidepressant treatment during pregnancy on obstetric outcomes among women previously treated for depression: an observational cohort study. *J Perinatol* 2017;**37**:1003–9. doi:10.1038/jp.2017.92

202 Shoaibi A, Fortin SP, Weinstein R, *et al.* Comparative Effectiveness of Famotidine in Hospitalized COVID-19 Patients. *Official journal of the American College of Gastroenterology | ACG* 9000;**Publish Ahead of Print**.https://journals.lww.com/ajg/Fulltext/9000/Comparative_Effectiveness_of_Famotidine_in.98891.aspx

203 Thurin NH, Lassalle R, Schuemie M, *et al.* Empirical assessment of case□based methods for drug safety alert identification in the French National Healthcare System database (SNDS): Methodology of the ALCAPONE project. *Pharmacoepidemiol Drug Saf* 2020;**29**:993–1000. doi:10.1002/pds.4983

204 Schuemie MJ, Ryan PB, DuMouchel W, *et al.* Interpreting observational studies: why empirical calibration is needed to correct *p* -values. *Statistics in Medicine* 2014;**33**:209–18. doi:10.1002/sim.5925

205 Neugebauer R, Schmittdiel JA, Zhu Z, *et al.* High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions. *Statist Med* 2015;**34**:753–81. doi:10.1002/sim.6377

206 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

207 Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International Journal of Epidemiology* 2018;**47**:2005–14. doi:10.1093/ije/dyy120

208 Weinstein RB, Ryan P, Berlin JA, *et al.* Channeling in the Use of Nonprescription Paracetamol and Ibuprofen in an Electronic Medical Records Database: Evidence and Implications. *Drug Saf* 2017;**40**:1279–92. doi:10.1007/s40264-017-0581-7

209 Zhang L, Wang Y, Schuemie MJ, *et al.* Adjusting for indirectly measured confounding using large-scale propensity score. *Journal of Biomedical Informatics* 2022;**134**:104204. doi:10.1016/j.jbi.2022.104204

210 Smith SM, Desai RA, Walsh MG, *et al.* Angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, and COVID-19-related outcomes: A patient-level analysis of the PCORnet blood pressure control lab. *American Heart Journal Plus: Cardiology Research and Practice* 2022;**13**:100112. doi:10.1016/j.ahjo.2022.100112

211 Hutchings A, O'Neill S, Lugo□Palacios D, *et al.* Effectiveness of emergency surgery for five common acute conditions: an instrumental variable analysis of a national routine database. *Anaesthesia* 2022;**77**:865–81. doi:10.1111/anae.15730

212 Fortin SP, Johnston SS, Schuemie MJ. Applied comparison of large-scale propensity score matching and cardinality matching for causal inference in observational research. *BMC Med Res Methodol* 2021;**21**:174. doi:10.1186/s12874-021-01365-z

213 Htoo PT, Measer G, Orr R, *et al.* Evaluating Confounding Control in Estimations of Influenza Antiviral Effectiveness in Electronic Health Plan Data. *American Journal of Epidemiology* 2022;**191**:908–20. doi:10.1093/aje/kwac020

214 Walker A, Patrick, Lauer, *et al.* A tool for assessing the feasibility of comparative effectiveness research. *CER* 2013;:11. doi:10.2147/CER.S40357

215 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. *Journal of the American Medical Informatics Association* Published Online First: 22 August 2020. doi:10.1093/jamia/ocaa124

216 Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD: LGPS AND LEOPARD. *Pharmacoepidem Drug Safe* 2011;**20**:292–9. doi:10.1002/pds.2051

217 EU-ADR consortium, Ferrajolo C, Coloma PM, *et al.* Signal Detection of Potentially Drug-Induced Acute Liver Injury in Children Using a Multi-Country Healthcare Database Network. *Drug Saf* 2014;**37**:99–108. doi:10.1007/s40264-013-0132-9

218 Meeker D, Jiang X, Matheny ME, *et al.* A system to build distributed multivariate models and manage disparate data sharing policies: implementation in the scalable national network for effectiveness research. *Journal of the American Medical Informatics Association* 2015;**22**:1187–95. doi:10.1093/jamia/ocv017

219 Lamer A, Depas N, Doutreligne M, *et al.* Erratum: Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. *Appl Clin Inform* 2020;**11**:e1–e1. doi:10.1055/s-0040-1702166

220 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

221 Quiroz JC, Chard T, Sa Z, *et al.* Extract, transform, load framework for the conversion of health databases to OMOP. *PLoS ONE* 2022;**17**:e0266911. doi:10.1371/journal.pone.0266911

222 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. *Healthcare informatics research* 2016;**22**:54–8.

223 Choi S, Choi SJ, Kim JK, *et al.* Preliminary feasibility assessment of CDM-based active surveillance using current status of medical device data in medical records and OMOP-CDM. *Sci Rep* 2021;**11**:24070. doi:10.1038/s41598-021-03332-6

224 Tan HX, Teo DCH, Lee D, *et al.* Applying the OMOP Common Data Model to Facilitate Benefit-Risk Assessments of Medicinal Products Using Real-World Data from Singapore and South Korea. *Healthc Inform Res* 2022;**28**:112–22. doi:10.4258/hir.2022.28.2.112

225 Klann JG, Joss MAH, Embree K, *et al.* Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PLoS ONE* 2019;**14**:e0212463. doi:10.1371/journal.pone.0212463

226 Si Y, Weng C. An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria. *Stud Health Technol Inform* 2017;**245**:950–4.

227 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**:119–34. doi:10.1007/s40264-012-0009-3

228 Ji H, Kim S, Yi S, *et al.* Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. *Journal of Biomedical Informatics* 2020;**107**:103459. doi:10.1016/j.jbi.2020.103459

229 Papez V, Moinat M, Payralbe S, *et al.* Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. *JAMIA Open* 2021;**4**:ooab001. doi:10.1093/jamiaopen/ooab001

230 Sathappan SMK, Jeon YS, Dang TK, *et al.* Transformation of Electronic Health Records and Questionnaire Data to OMOP CDM: A Feasibility Study Using SG_T2DM Dataset. *Appl Clin Inform* 2021;**12**:757–67. doi:10.1055/s-0041-1732301

231 Matcho A, Ryan P, Fife D, *et al.* Fidelity Assessment of a Clinical Practice Research Datalink Conversion to the OMOP Common Data Model. *Drug Saf* 2014;**37**:945–59. doi:10.1007/s40264-014-0214-3

232 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. *Journal of Biomedical Informatics* 2022;**127**:104002. doi:10.1016/j.jbi.2022.104002

233 Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc* 1993;**81**:170–7.

234 Noy NF, Shah NH, Whetzel PL, *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* 2009;**37**:W170–3. doi:10.1093/nar/gkp440

235 Ostropolets A, Dymshyts D, Reich C, *et al.* Comprehensive Procedure Hierarchy. 2018.

236 Dymshyts D, Ostropolets A, Reich C. International RxNorm Extension to support the expansion of the OHDSI Network beyond the US. 2017.https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:rxe_poster_2017.pdf

237 Ostropolets A. OHDSI RxNorm Extension: 5 major drug markets in RxNorm notation, comparison and lessons learned. 2017.

238 Shivade C, Raghavan P, Fosler-Lussier E, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association* 2014;**21**:221–30. doi:10.1136/amiajnl-2013-001935

239 Vlasschaert MEO, Bejaimal SAD, Hackam DG, *et al.* Validity of Administrative Database Coding for Kidney Disease: A Systematic Review. *American Journal of Kidney Diseases* 2011;**57**:29–43. doi:10.1053/j.ajkd.2010.08.031

240 Gentile G, Postorino M, Mooring RD, *et al.* Estimated GFR reporting is not sufficient to allow detection of chronic kidney disease in an Italian regional hospital. *BMC Nephrology* 2009;**10**. doi:10.1186/1471-2369-10-24

241 Kern EFO, Maney M, Miller DR, *et al.* Failure of ICD-9-CM Codes to Identify Patients with Comorbid Chronic Kidney Disease in Diabetes. *Health Services Research* 2006;**41**:564–80. doi:10.1111/j.1475-6773.2005.00482.x

242 Ostropolets, Anna A. Investigating concept heterogeneity and granularity in the OHDSI network. 2020.https://www.ohdsi.org/wp-content/uploads/2020/10/Ostropolets_Plenary.pdf (accessed 28 Oct 2020).

243 Rowland D, Lyons B. Medicare, Medicaid, and the elderly poor. *Health Care Financ Rev* 1996;**18**:61–85.

244 Hruschka DJ, Schwartz D, St.John DC, *et al.* Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods* 2004;**16**:307–31. doi:10.1177/1525822X04266540

245 Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *Journal of the American Medical Informatics Association* 2011;**18**:432–40. doi:10.1136/amiajnl-2010-000045

246 Cui L, Bodenreider O, Shi J, *et al.* Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs. *Journal of Biomedical Informatics* 2018;**78**:177–84. doi:10.1016/j.jbi.2017.12.010

247 Ostropolets A. Phenotype algorithm and data source reporting in top clinical journals: where we are and where should we go? 2020. https://www.ohdsi.org/wp-content/uploads/2020/10/Anna-Ostropolets-Ostropolets-db_provenane-Abstract.pdf

248 Lanes S, Brown JS, Haynes K, *et al.* Identifying health outcomes in healthcare databases: Identifying Health Outcomes. *Pharmacoepidemiol Drug Saf* 2015;**24**:1009–16. doi:10.1002/pds.3856

249 Chen P, Dong W, Lu X, *et al.* Deep representation learning for individualized treatment effect estimation using electronic health records. *Journal of biomedical informatics* 2019;**100**:103303.

250 Gatto NM, Campbell UB, Rubinstein E, *et al.* The Structured Process to Identify Fit□For□ Purpose Data: A Data Feasibility Assessment Framework. *Clin Pharma and Therapeutics* 2022;**111**:122–34. doi:10.1002/cpt.2466

251 Navarro G. A guided tour to approximate string matching. *ACM Comput Surv* 2001;**33**:31–88. doi:10.1145/375360.375365

252 Kagawa R, Kawazoe Y, Ida Y, *et al.* Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach. *J Diabetes Sci Technol* 2017;**11**:791–9. doi:10.1177/1932296816681584

253 Spratt SE, Pereira K, Granger BB, *et al.* Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *Journal of the American Medical Informatics Association* 2017;**24**:e121–8. doi:10.1093/jamia/ocw123

254 Blecker S, Katz SD, Horwitz LI, *et al.* Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. *JAMA Cardiol* 2016;**1**:1014. doi:10.1001/jamacardio.2016.3236

255 Slaby I, Hain HS, Abrams D, *et al.* An electronic health record (EHR) phenotype algorithm to identify patients with attention deficit hyperactivity disorders (ADHD) and psychiatric comorbidities. *J Neurodevelop Disord* 2022;**14**:37. doi:10.1186/s11689-022-09447-9

256 Prieto-Alhambra D, Kostka K, Duarte-Salles T, *et al.* Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. 2021.

257 Recalde M, Roel E, Pistillo A, *et al.* Characteristics and outcomes of 627 044 COVID-19 patients with and without obesity in the United States, Spain, and the United Kingdom. Published Online First: 3 September 2020. doi:10.1101/2020.09.02.20185173

258 Ostropolets A, Li X, Makadia R, *et al.* Empirical evaluation of the sensitivity of background incidence rate characterization for adverse events across an international observational data network. Health Informatics 2021. doi:10.1101/2021.06.27.21258701

259 Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences* 2016;**113**:7329–36. doi:10.1073/pnas.1510502113

260 Brown JS, Maro JC, Nguyen M, *et al.* Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the Food and Drug Administration's Sentinel system. *Journal of the American Medical Informatics Association* 2020;**27**:793–7. doi:10.1093/jamia/ocaa028

261 Overby CL, Pathak J, Gottesman O, *et al.* A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. *J Am Med Inform Assoc* 2013;**20**:e243–52. doi:10.1136/amiajnl-2013-001930

262 Seymour CW, Kennedy JN, Wang S, *et al.* Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *JAMA* 2019;**321**:2003. doi:10.1001/jama.2019.5791

263 Albogami Y, Cusi K, Daniels MJ, *et al.* Glucagon-Like Peptide 1 Receptor Agonists and Chronic Lower Respiratory Disease Exacerbations Among Patients With Type 2 Diabetes. *Dia Care* 2021;**44**:1344–52. doi:10.2337/dc20-1794

264 ATLAS. 2022.https://github.com/OHDSI/Atlas (accessed 23 Apr 2022).

265 GitHub - aostropolets/ReproducibilityChallenge. https://github.com/aostropolets/ReproducibilityChallenge (accessed 17 Sep 2022).

266 Fletcher S, Islam MZ. Comparing sets of patterns with the Jaccard index. *AJIS* 2018;**22**. doi:10.3127/ajis.v22i0.1538

267 Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics - Simulation and Computation* 2009;**38**:1228–34. doi:10.1080/03610910902859574

268 Andrade SE, Raebel MA, Morse AN, *et al.* Use of prescription medications with a potential for fetal harm among pregnant women. *Pharmacoepidem Drug Safe* 2006;**15**:546–54. doi:10.1002/pds.1235

269 Matcho A, Ryan P, Fife D, *et al.* Inferring pregnancy episodes and outcomes within a network of observational databases. *PLoS ONE* 2018;**13**:e0192033. doi:10.1371/journal.pone.0192033

270 Wei W-Q, Leibson CL, Ransom JE, *et al.* The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *International Journal of Medical Informatics* 2013;**82**:239–47. doi:10.1016/j.ijmedinf.2012.05.015

271 Wang L, Olson JE, Bielinski SJ, *et al.* Impact of Diverse Data Sources on Computational Phenotyping. *Front Genet* 2020;**11**:556. doi:10.3389/fgene.2020.00556

272 Coresh J, Astor BC, McQuillan G, *et al.* Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate. *American Journal of Kidney Diseases* 2002;**39**:920–9. doi:10.1053/ajkd.2002.32765

273 Fleet JL, Dixon SN, Shariff SZ, *et al.* Detecting chronic kidney disease in population-based administrative databases using an algorithm of hospital encounter and physician claim codes. *BMC Nephrology* 2013;**14**. doi:10.1186/1471-2369-14-81

274 Ronksley PE, Tonelli M, Quan H, *et al.* Validating a case definition for chronic kidney disease using administrative data. *Nephrology Dialysis Transplantation* 2012;**27**:1826–31. doi:10.1093/ndt/gfr598

275 Stevens LA, Fares G, Fleming J, *et al.* Low rates of testing and diagnostic codes usage in a commercial clinical laboratory: evidence for lack of physician awareness of chronic kidney disease. *J Am Soc Nephrol* 2005;**16**:2439–48. doi:10.1681/ASN.2005020192

276 Winkelmayer WC, Schneeweiss S, Mogun H, *et al.* Identification of Individuals With CKD From Medicare Claims Data: A Validation Study. *American Journal of Kidney Diseases* 2005;**46**:225–32. doi:10.1053/j.ajkd.2005.04.029

277 Grams ME, Plantinga LC, Hedgeman E, *et al.* Validation of CKD and Related Conditions in Existing Data Sets: A Systematic Review. *American Journal of Kidney Diseases* 2011;**57**:44–54. doi:10.1053/j.ajkd.2010.05.013

278 Shang N, Khan A, Polubriaginof F, *et al.* Medical records-based chronic kidney disease phenotype for clinical care and “big data” observational and genetic studies. *npj Digit Med* 2021;**4**:70. doi:10.1038/s41746-021-00428-1

279 Levey AS, de Jong PE, Coresh J, *et al.* The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. *Kidney International* 2011;**80**:17–28. doi:10.1038/ki.2010.483

280 Ostropolets A. Investigating portability of EHR-derinved pehnotypes. 2022.https://github.com/aostropolets/ChronicKidneyDisorderPhenotyping (accessed 14 Sep 2022).

281 THEMIS - The Answer to Building Consistent Common Data Models Across Organizations. 2018.

282 KDIGO 2017 Clinical Practice Guideline Update for the Diagnosis, Evaluation, Prevention, and Treatment of Chronic Kidney Disease–Mineral and Bone Disorder (CKD-MBD). *Kidney International Supplements* 2017;**7**:1–59. doi:10.1016/j.kisu.2017.04.001

283 Nattinger AB, Laud PW, Bajorunaite R, *et al.* An Algorithm for the Use of Medicare Claims Data to Identify Women with Incident Breast Cancer. *Health Serv Res* 2004;**39**:1733–50. doi:10.1111/j.1475-6773.2004.00315.x

284 Anna Ostropolets M, RuiJun Chen M, Matthew Spotnitz M, *et al.* Phenotype algorithm and data source reporting in top clinical journals: where we are and where should we go? 2020.

285 Rubbo B, Fitzpatrick NK, Denaxas S, *et al.* Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *International Journal of Cardiology* 2015;**187**:705–11. doi:10.1016/j.ijcard.2015.03.075

286 Jensen PN, Johnson K, Floyd J, *et al.* A systematic review of validated methods for identifying atrial fibrillation using administrative data: DETECTION OF ATRIAL FIBRILLATION IN CLAIMS. *Pharmacoepidemiol Drug Saf* 2012;**21**:141–7. doi:10.1002/pds.2317

287 Faust E, Pivneva I, Yang K, *et al.* Real-World Treatment Profiles, Clinical Outcomes, and Healthcare Resource Utilization of Patients with Migraine Prescribed Erenumab: A Multicenter Chart-Review Study of US Headache Centers. *Neurol Ther* 2021;**10**:293–306. doi:10.1007/s40120-021-00245-4

288 Helwig U, Mross M, Schubert S, *et al.* Real-world clinical effectiveness and safety of vedolizumab and anti-tumor necrosis factor alpha treatment in ulcerative colitis and Crohn's disease patients: a German retrospective chart review. *BMC Gastroenterol* 2020;**20**:211. doi:10.1186/s12876-020-01332-w

289 Mohty M, Cavo M, Fink L, *et al.* Understanding mortality in multiple myeloma: Findings of a European retrospective chart review. *Eur J Haematol* 2019;**103**:107–15. doi:10.1111/ejh.13264

290 Kho ME, Duffett M, Willison DJ, *et al.* Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 2009;**338**:b866–b866. doi:10.1136/bmj.b866

291 Yin Z, Tong J, Chen Y, *et al.* A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data. *Journal of the American Medical Informatics Association* 2021;**29**:52–61. doi:10.1093/jamia/ocab222

292 Liu X, Chubak J, Hubbard RA, *et al.* SAT: a Surrogate-Assisted Two-wave case boosting sampling method, with application to EHR-based association studies. *Journal of the American Medical Informatics Association* 2022;**29**:918–27. doi:10.1093/jamia/ocab267

293 Garza MY, Ounpraseuth S, Hu Z, *et al.* Measuring and Controlling Medical Record Abstraction (MRA) Error Rates in an Observational Study. In Review 2022. doi:10.21203/rs.3.rs-1225727/v1

294 Zozus MN, Pieper C, Johnson CM, *et al.* Factors Affecting Accuracy of Data Abstracted from Medical Records. *PLoS ONE* 2015;**10**:e0138649. doi:10.1371/journal.pone.0138649

295 Widdifield J, Labrecque J, Lix L, *et al.* Systematic Review and Critical Appraisal of Validation Studies to Identify Rheumatic Diseases in Health Administrative Databases: Review of Administrative Data Validation Studies. *Arthritis Care & Research* 2013;**65**:1490–503. doi:10.1002/acr.21993

296 Bayley KB, Belnap T, Savitz L, *et al.* Challenges in Using Electronic Health Record Data for CER: Experience of 4 Learning Organizations and Solutions Applied. *Medical Care* 2013;**51**:S80–6.

297 Hardjojo A, Gunachandran A, Pang L, *et al.* Validation of a Natural Language Processing Algorithm for Detecting Infectious Disease Symptoms in Primary Care Electronic Medical Records in Singapore. *JMIR Med Inform* 2018;**6**:e36. doi:10.2196/medinform.8204

298 Zhou L, Suominen H, Gedeon T. Adapting State-of-the-Art Deep Language Models to Clinical Information Extraction Systems: Potentials, Challenges, and Solutions. *JMIR Med Inform* 2019;**7**:e11499. doi:10.2196/11499

299 Brunekreef TE, Otten HG, Bosch SC, *et al.* Text Mining of Electronic Health Records Can Accurately Identify and Characterize Patients With Systemic Lupus Erythematosus. *ACR Open Rheuma* 2021;**3**:65–71. doi:10.1002/acr2.11211

300 Jorge A, Castro VM, Barnado A, *et al.* Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. *Seminars in Arthritis and Rheumatism* 2019;**49**:84–90. doi:10.1016/j.semarthrit.2019.01.002

301 McKenzie J, Rajapakshe R, Shen H, *et al.* A Semiautomated Chart Review for Assessing the Development of Radiation Pneumonitis Using Natural Language Processing: Diagnostic Accuracy and Feasibility Study. *JMIR Med Inform* 2021;**9**:e29241. doi:10.2196/29241

302 Afzal N, Sohn S, Abram S, *et al.* Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of Vascular Surgery* 2017;**65**:1753–61. doi:10.1016/j.jvs.2016.11.031

303 Ford E, Carroll JA, Smith HE, *et al.* Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association* 2016;**23**:1007–15. doi:10.1093/jamia/ocv180

304 Lamy M, Pereira R, Ferreira JC, *et al.* Extracting Clinical Information from Electronic Medical Records. In: Novais P, Jung JJ, Villarrubia González G, *et al.*, eds. *Ambient Intelligence – Software and Applications –, 9th International Symposium on Ambient Intelligence*. Cham: : Springer International Publishing 2019. 113–20. doi:10.1007/978-3-030-01746-0_13

305 Lovis C, Baud RH, Planche P. Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics* 2000;**58–59**:101–10. doi:10.1016/S1386-5056(00)00079-4

306 Bickley LS. *Bates' guide to physical examination and history taking*. Eighth edition / Lynn S. Bickley, Peter G. Szilagyi. Philadelphia : Lippincott Williams & Wilkins, [2003] ©2003 2003. https://search.library.wisc.edu/catalog/999931788202121

307 Liu S, Wang L, Ihrke D, *et al.* Correlating Lab Test Results in Clinical Notes with Structured Lab Data: A Case Study in HbA1c and Glucose. *AMIA Jt Summits Transl Sci Proc* 2017;**2017**:221–8.

308 Silen W, Cope Z. *Cope's early diagnosis of the acute abdomen*. 22nd ed. / revised by William Silen. Oxford ; New York: : Oxford University Press 2010.

309 Addiss DG, Shaffer N, Fowler BS, *et al.* THE EPIDEMIOLOGY OF APPENDICITIS AND APPENDECTOMY IN THE UNITED STATES. *American Journal of Epidemiology* 1990;**132**:910–25. doi:10.1093/oxfordjournals.aje.a115734

310 Appendicitis | PheKB. https://phekb.org/phenotype/appendicitis (accessed 31 Oct 2022).

311 Columbia COPD Implementation | PheKB. https://phekb.org/implementation/columbia-copd-implementation (accessed 31 Oct 2022).

312 Chronic Kidney Disease | PheKB. https://phekb.org/phenotype/chronic-kidney-disease (accessed 31 Oct 2022).

313 Type 1 Diabetes | PheKB. https://phekb.org/phenotype/type-1-diabetes (accessed 31 Oct 2022).

314 Lim C-Y, In J. Considerations for crossover design in clinical study. *Korean J Anesthesiol* 2021;**74**:293–9. doi:10.4097/kja.21165

315 Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *TQMP* 2012;**8**:23–34. doi:10.20982/tqmp.08.1.p023

316 Agresti A. *Categorical data analysis*. New York: : Wiley 1990.

317 Zhang D, Yin C, Zeng J, *et al.* Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak* 2020;**20**:280. doi:10.1186/s12911-020-01297-6

318 Kaur H, Sohn S, Wi C-I, *et al.* Automated chart review utilizing natural language processing algorithm for asthma predictive index. *BMC Pulm Med* 2018;**18**:34. doi:10.1186/s12890-018-0593-9

319 Ford E, Nicholson A, Koeling R, *et al.* Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? *BMC Med Res Methodol* 2013;**13**:105. doi:10.1186/1471-2288-13-105

320 Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. *JAMA Netw Open* 2022;**5**:e2233348. doi:10.1001/jamanetworkopen.2022.33348

321 Vassar M, Matthew H. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof* 2013;**10**:12. doi:10.3352/jeehp.2013.10.12

322 Liddy C, Wiens M, Hogg W. Methods to Achieve High Interrater Reliability in Data Collection From Primary Care Medical Records. *The Annals of Family Medicine* 2011;**9**:57–62. doi:10.1370/afm.1195

323 Tayefi M, Ngo P, Chomutare T, *et al.* Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comp Stat* 2021;**13**. doi:10.1002/wics.1549

324 Wiese AD, Roumie CL, Buse JB, *et al.* Performance of a computable phenotype for identification of patients with diabetes within PCORnet: The Patient□Centered Clinical Research Network. *Pharmacoepidemiol Drug Saf* 2019;**28**:632–9. doi:10.1002/pds.4718

325 Newton KM, Peissig PL, Kho AN, *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association* 2013;**20**:e147–54. doi:10.1136/amiajnl-2012-000896

326 Wenderfer SE, Chang JC, Goodwin Davies A, *et al.* Using a Multi-Institutional Pediatric Learning Health System to Identify Systemic Lupus Erythematosus and Lupus Nephritis:

Development and Validation of Computable Phenotypes. *CJASN* 2022;**17**:65–74. doi:10.2215/CJN.07810621

327 Humphries KH, Rankin JM, Carere RG, *et al.* Co-morbidity data in outcomes research Are clinical data derived from administrative databases a reliable alternative to chart review? *Journal of Clinical Epidemiology* 2000;**53**:343–9. doi:10.1016/S0895-4356(99)00188-2

328 Dixon BE. Public Health: Interoperability Applications to Support Population Health. In: Hübner UH, Mustata Wilson G, Morawski TS, *et al.*, eds. *Nursing Informatics*. Cham: : Springer International Publishing 2022. 339–54. doi:10.1007/978-3-030-91237-6_23

329 Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association* 2015;**22**:938–47. doi:10.1093/jamia/ocv032

330 Chen IY, Joshi S, Ghassemi M, *et al.* Probabilistic Machine Learning for Healthcare. *Annu Rev Biomed Data Sci* 2021;**4**:393–415. doi:10.1146/annurev-biodatasci-092820-033938

331 Kulmanov M, Smaili FZ, Gao X, *et al.* Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics* 2021;**22**:bbaa199. doi:10.1093/bib/bbaa199

332 King AJ, Cooper GF, Clermont G, *et al.* Using machine learning to selectively highlight patient information. *Journal of Biomedical Informatics* 2019;**100**:103327. doi:10.1016/j.jbi.2019.103327

333 Tajgardoon M, Cooper GF, King AJ, *et al.* Modeling physician variability to prioritize relevant medical record information. *JAMIA Open* 2021;**3**:602–10. doi:10.1093/jamiaopen/ooaa058

334 Lenert MC, Walsh CG, Miller RA. Discovering hidden knowledge through auditing clinical diagnostic knowledge bases. *Journal of Biomedical Informatics* 2018;**84**:75–81. doi:10.1016/j.jbi.2018.06.014

335 Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* 2011;**46**:399–424. doi:10.1080/00273171.2011.568786

336 Mahaux O, Bauchau V, Van Holle L. Pharmacoepidemiological considerations in observed□to□expected analyses for vaccines. *Pharmacoepidemiol Drug Saf* 2016;**25**:215–22. doi:10.1002/pds.3918

337 Grosso A, Douglas I, MacAllister R, *et al.* Use of the self-controlled case series method in drug safety assessment. *Expert Opinion on Drug Safety* 2011;**10**:337–40. doi:10.1517/14740338.2011.562187

338 Black S, Eskola J, Siegrist C-A, *et al.* Importance of background rates of disease in assessment of vaccine safety during mass immunisation with pandemic H1N1 influenza vaccines. *The Lancet* 2009;**374**:2115–22. doi:10.1016/S0140-6736(09)61877-8

339 Lao KSJ, Chui CSL, Man KKC, *et al.* Medication safety research by observational study design. *Int J Clin Pharm* Published Online First: 22 March 2016. doi:10.1007/s11096-016-0285-6

340 Trifirò G, Crisafulli S. A New Era of Pharmacovigilance: Future Challenges and Opportunities. *Front Drug Saf Regul* 2022;**2**:866898. doi:10.3389/fdsfr.2022.866898

341 Banda JM, Evans L, Vanguri RS, *et al.* A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data* 2016;**3**:160026. doi:10.1038/sdata.2016.26

342 Voss EA, Boyce RD, Ryan PB, *et al.* Accuracy of an automated knowledge base for identifying drug adverse reactions. *Journal of Biomedical Informatics* 2017;**66**:72–81. doi:10.1016/j.jbi.2016.12.005

343 Hoffman KB, Dimbil M, Erdman CB, *et al.* The Weber Effect and the United States Food and Drug Administration's Adverse Event Reporting System (FAERS): Analysis of Sixty-Two Drugs Approved from 2006 to 2010. *Drug Saf* 2014;**37**:283–94. doi:10.1007/s40264-014-0150-2

344 Rahman MdM, Alatawi Y, Cheng N, *et al.* Methodological Considerations for Comparison of Brand Versus Generic Versus Authorized Generic Adverse Event Reports in the US Food and Drug Administration Adverse Event Reporting System (FAERS). *Clin Drug Investig* 2017;**37**:1143–52. doi:10.1007/s40261-017-0574-4

345 Cai R, Liu M, Hu Y, *et al.* Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. *Artificial Intelligence in Medicine* 2017;**76**:7–15. doi:10.1016/j.artmed.2017.01.004

346 Harpaz R, DuMouchel W, LePendu P, *et al.* Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System. *Clin Pharmacol Ther* 2013;**93**:539–46. doi:10.1038/clpt.2013.24

347 Liu X, Chen H. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums. In: Zeng D, Yang CC, Tseng VS, *et al.*, eds. *Smart Health*. Berlin, Heidelberg: : Springer Berlin Heidelberg 2013. 134–50. doi:10.1007/978-3-642-39844-5_16

348 Smith K, Golder S, Sarker A, *et al.* Methods to Compare Adverse Events in Twitter to FAERS, Drug Information Databases, and Systematic Reviews: Proof of Concept with Adalimumab. *Drug Saf* 2018;**41**:1397–410. doi:10.1007/s40264-018-0707-6

349 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

350 Food and Drug Administration, Center for Biologics Evaluation and Research (CBER) Biologics Effectiveness and Safety (BEST) Initiative. Background Rates of Adverse Events of Special Interest for COVID-19 Vaccine Safety Monitoring. 2020;:32.

351 European Network of Centres for Pharmacoepidemiology and Phamacovigilance. ACCESS. Background rates of Adverse Events of Special Interest for monitoring COVID-19 vaccines. 2020;:54.

352 Black SB, Law B, Chen RT, *et al.* The critical role of background rates of possible adverse events in the assessment of COVID-19 vaccine safety. *Vaccine* 2021;**39**:2712–8. doi:10.1016/j.vaccine.2021.03.016

353 Nasreen S, Calzavara A, Sundaram M, *et al.* Background rates of hospitalizations and emergency department visits for selected thromboembolic and coagulation disorders in Ontario, Canada, 2015 to 2020, to inform COVID-19 vaccine safety surveillance. Public and Global Health 2021. doi:10.1101/2021.04.02.21254856

354 Burn E, Li X, Kostka K, *et al.* Background rates of five thrombosis with thrombocytopenia syndromes of special interest for COVID-19 vaccine safety surveillance: incidence between 2017 and 2019 and patient profiles from 20.6 million people in six European countries. Hematology 2021. doi:10.1101/2021.05.12.21257083

355 Umasunthar T, Leonardi□Bee J, Hodes M, *et al.* Incidence of fatal food anaphylaxis in people with food allergy: a systematic review and meta□analysis. *Clin Exp Allergy* 2013;**43**:1333–41. doi:10.1111/cea.12211

356 Susantitaphong P, Cruz DN, Cerda J, *et al.* World Incidence of AKI: A Meta-Analysis. *CJASN* 2013;**8**:1482–93. doi:10.2215/CJN.00710113

357 Hirsch L, Jette N, Frolkis A, *et al.* The Incidence of Parkinson's Disease: A Systematic Review and Meta-Analysis. *Neuroepidemiology* 2016;**46**:292–300. doi:10.1159/000445751

358 Dasgupta P, Henshaw C, Youlden DR, *et al.* Global Trends in Incidence Rates of Primary Adult Liver Cancers: A Systematic Review and Meta-Analysis. *Front Oncol* 2020;**10**:171. doi:10.3389/fonc.2020.00171

359 Dodd CN, de Ridder M, Huang W-T, *et al.* Incidence rates of narcolepsy diagnoses in Taiwan, Canada, and Europe: The use of statistical simulation to evaluate methods for the rapid assessment of potential safety issues on a population level in the SOMNIA study. *PLoS ONE* 2018;**13**:e0204799. doi:10.1371/journal.pone.0204799

360 Lin KJ, García Rodríguez LA, Hernández-Díaz S. Systematic review of peptic ulcer disease incidence rates: do studies without validation provide reliable estimates?: INCIDENCE OF PEPTIC ULCER. *Pharmacoepidemiol Drug Saf* 2011;**20**:718–28. doi:10.1002/pds.2153

361 Sejvar JJ, Baughman AL, Wise M, *et al.* Population Incidence of Guillain-Barré Syndrome: A Systematic Review and Meta-Analysis. *Neuroepidemiology* 2011;**36**:123–33. doi:10.1159/000324710

362 Willame C, Dodd C, van der Aa L, *et al.* Incidence Rates of Autoimmune Diseases in European Healthcare Databases: A Contribution of the ADVANCE Project. *Drug Saf* 2021;**44**:383–95. doi:10.1007/s40264-020-01031-1

363 Hanratty B. Sex differences in risk factors, treatment and mortality after acute myocardial infarction: an observational study. *Journal of Epidemiology & Community Health* 2000;**54**:912–6. doi:10.1136/jech.54.12.912

364 Gracia Gutiérrez A, Poblador-Plou B, Prados-Torres A, *et al.* Sex Differences in Comorbidity, Therapy, and Health Services' Use of Heart Failure in Spain: Evidence from Real-World Data. *Int J Environ Res Public Health* 2020;**17**. doi:10.3390/ijerph17062136

365 Linn FHH, Rinkel GJE, Algra A, *et al.* Incidence of Subarachnoid Hemorrhage: Role of Region, Year, and Rate of Computed Tomography: A Meta-Analysis. *Stroke* 1996;**27**:625–9. doi:10.1161/01.STR.27.4.625

366 Kanaya AM, Adler N, Moffet HH, *et al.* Heterogeneity of Diabetes Outcomes Among Asians and Pacific Islanders in the U.S.: The Diabetes Study of Northern California (DISTANCE). *Diabetes Care* 2011;**34**:930–7. doi:10.2337/dc10-1964

367 Huang RJ, Sharp N, Talamoa RO, *et al.* One Size Does Not Fit All: Marked Heterogeneity in Incidence of and Survival from Gastric Cancer among Asian American Subgroups. *Cancer Epidemiol Biomarkers Prev* 2020;**29**:903–9. doi:10.1158/1055-9965.EPI-19-1482

368 Idrees R, Fatima S, Abdul-Ghafar J, *et al.* Cancer prevalence in Pakistan: meta-analysis of various published studies to determine variation in cancer figures resulting from marked population heterogeneity in different parts of the country. *World J Surg Onc* 2018;**16**:129. doi:10.1186/s12957-018-1429-z

369 Marty L, Cazein F, Panjo H, *et al.* Revealing geographical and population heterogeneity in HIV incidence, undiagnosed HIV prevalence and time to diagnosis to improve prevention and care: estimates for France. *J Intern AIDS Soc* 2018;**21**:e25100. doi:10.1002/jia2.25100

370 Beghi E, Chiò A, Couratier P, *et al.* The epidemiology and treatment of ALS: Focus on the heterogeneity of the disease and critical appraisal of therapeutic trials. *Amyotrophic Lateral Sclerosis* 2011;**12**:1–10. doi:10.3109/17482968.2010.502940

371 Cologne J, Kim J, Sugiyama H, *et al.* Effect of Heterogeneity in Background Incidence on Inference about the Solid-Cancer Radiation Dose Response in Atomic Bomb Survivors. *Radiation Research* 2019;**192**:388. doi:10.1667/RR15127.1

372 PHOEBE. https://data.ohdsi.org/PHOEBE/ (accessed 6 Feb 2022).

373 CohortDiagnostics. 2022.https://github.com/OHDSI/CohortDiagnostics (accessed 6 Feb 2022).

374 Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Soft* 2010;**36**. doi:10.18637/jss.v036.i03

375 Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, *et al.* Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods* 2006;**11**:193–206. doi:10.1037/1082-989X.11.2.193

376 Covid-19 Vaccine AESI Incidence Characterization protocol. GitHub. https://github.com/ohdsi-studies/Covid19VaccineAesiIncidenceCharacterization (accessed 4 May 2021).

377 Hense S, Schink T, Kreisel SH, *et al.* Estimation of Background Incidence Rates of Guillain-Barré Syndrome in Germany - A Retrospective Cohort Study with Electronic Healthcare Data. *Neuroepidemiology* 2014;**43**:244–52. doi:10.1159/000369344

378 Koopman C, Bots ML, van Oeffelen AAM, *et al.* Population trends and inequalities in incidence and short-term outcome of acute myocardial infarction between 1998 and 2007. *International Journal of Cardiology* 2013;**168**:993–8. doi:10.1016/j.ijcard.2012.10.036

379 Barker-Collo S, Bennett DA, Krishnamurthi RV, *et al.* Sex Differences in Stroke Incidence, Prevalence, Mortality and Disability-Adjusted Life Years: Results from the Global Burden of Disease Study 2013. *Neuroepidemiology* 2015;**45**:203–14. doi:10.1159/000441103

380 Fairweather D, Cooper LT, Blauwet LA. Sex and Gender Differences in Myocarditis and Dilated Cardiomyopathy. *Current Problems in Cardiology* 2013;**38**:7–46. doi:10.1016/j.cpcardiol.2012.07.003

381 Wang Y, Allen KJ, Suaini NHA, *et al.* The global incidence and prevalence of anaphylaxis in children in the general population: A systematic review. *Allergy* 2019;**74**:1063–80. doi:10.1111/all.13732

382 Drozd EM, Miller L, Johnsrud M. Impact of Pharmacist Immunization Authority on Seasonal Influenza Immunization Rates Across States. *Clinical Therapeutics* 2017;**39**:1563-1580.e17. doi:10.1016/j.clinthera.2017.07.004

383 Marrero O, Hung EY, Hauben M. Seasonal and Geographic Variation in Adverse Event Reporting. *Drugs - Real World Outcomes* 2016;**3**:297–306. doi:10.1007/s40801-016-0081-6

384 Nagarajan V, Fonarow GC, Ju C, *et al.* Seasonal and circadian variations of acute myocardial infarction: Findings from the Get With The Guidelines–Coronary Artery Disease (GWTG-CAD) program. *American Heart Journal* 2017;**189**:85–93. doi:10.1016/j.ahj.2017.04.002

385 Tschöpe C, Ammirati E, Bozkurt B, *et al.* Myocarditis and inflammatory cardiomyopathy: current evidence and future directions. *Nat Rev Cardiol* 2021;**18**:169–93. doi:10.1038/s41569-020-00435-x

386 Chaaban MR, Warren Z, Baillargeon JG, *et al.* Epidemiology and trends of anaphylaxis in the United States, 2004□2016. *Int Forum Allergy Rhinol* 2019;**9**:607–14. doi:10.1002/alr.22293

387 Kadambari S, Okike I, Ribeiro S, *et al.* Seven-fold increase in viral meningo-encephalitis reports in England and Wales during 2004–2013. *Journal of Infection* 2014;**69**:326–32. doi:10.1016/j.jinf.2014.05.012

388 Hamedani AG, Blank L, Thibault DP, *et al.* Impact of ICD-9 to ICD-10 Coding Transition on Prevalence Trends in Neurology. *Neurology: Clinical Practice* 2021;:10.1212/CPJ.0000000000001046. doi:10.1212/CPJ.0000000000001046

389 Katan M, Luft A. Global Burden of Stroke. Published Online First: 2018. doi:10.5167/UZH-159894

390 Lee JS, Kim YH. Epidemiological trends of Bell's palsy treated with steroids in Korea between 2008 and 2018. *Muscle & Nerve* 2021;:mus.27213. doi:10.1002/mus.27213

391 Zayet S, Kadiane-Oussou NJ, Lepiller Q, *et al.* Clinical features of COVID-19 and influenza: a comparative study on Nord Franche-Comte cluster. *Microbes and Infection* 2020;**22**:481–8. doi:10.1016/j.micinf.2020.05.016

392 Verity R, Okell LC, Dorigatti I, *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases* 2020;**20**:669–77. doi:10.1016/S1473-3099(20)30243-7

393 Dooling K, Marin M, Wallace M, *et al.* The Advisory Committee on Immunization Practices' Updated Interim Recommendation for Allocation of COVID-19 Vaccine — United States, December 2020. *MMWR Morb Mortal Wkly Rep* 2021;**69**:1657–60. doi:10.15585/mmwr.mm695152e2

394 Coronavirus (COVID-19) Vaccinations - Statistics and Research. Our World in Data. https://ourworldindata.org/covid-vaccinations (accessed 31 May 2021).

395 Lee BY, Brown ST, Korch GW, *et al.* A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 H1N1 influenza pandemic. *Vaccine* 2010;**28**:4875–9. doi:10.1016/j.vaccine.2010.05.002

396 Petrie JG, Ohmit SE, Cheng CK, *et al.* Influenza Vaccine Effectiveness Against Antigenically Drifted Influenza Higher Than Expected in Hospitalized Adults: 2014–2015. *Clin Infect Dis* 2016;**63**:1017–25. doi:10.1093/cid/ciw432

397 Pilishvili T, Fleming-Dutra KE, Farrar JL, *et al.* Interim estimates of vaccine effectiveness of Pfizer-BioNTech and Moderna COVID-19 vaccines among health care personnel—33 US sites, January–March 2021. 2021.

398 Ostropolets A. Effects of anchoring on baseline patient characteristics in patients vaccinated with influenza and COVID-19 vaccine. 2022.https://github.com/aostropolets/VaccineAnchoring (accessed 14 Sep 2022).

399 Sadoff J, Gray G, Vandebosch A, *et al.* Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *N Engl J Med* 2021;**384**:2187–201. doi:10.1056/NEJMoa2101544

400 Baden LR, El Sahly HM, Essink B, *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med* 2021;**384**:403–16. doi:10.1056/NEJMoa2035389

401 Polack FP, Thomas SJ, Kitchin N, *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med* 2020;**383**:2603–15. doi:10.1056/NEJMoa2034577

402 Voysey M, Clemens SAC, Madhi SA, *et al.* Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet* 2021;**397**:99–111. doi:10.1016/S0140-6736(20)32661-1

403 Thompson MG, Stenehjem E, Grannis S, *et al.* Effectiveness of Covid-19 Vaccines in Ambulatory and Inpatient Care Settings. *N Engl J Med* 2021;:NEJMoa2110362. doi:10.1056/NEJMoa2110362

404 Tartof SY, Slezak JM, Fischer H, *et al.* Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. *The Lancet* 2021;**398**:1407–16. doi:10.1016/S0140-6736(21)02183-8

405 Haas EJ, Angulo FJ, McLaughlin JM, *et al.* Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet* 2021;**397**:1819–29. doi:10.1016/S0140-6736(21)00947-8

406 Kissling E, Hooiveld M, Sandonis Martín V, *et al.* Vaccine effectiveness against symptomatic SARS-CoV-2 infection in adults aged 65 years and older in primary care: I-MOVE-COVID-19 project, Europe, December 2020 to May 2021. *Eurosurveillance* 2021;**26**. doi:10.2807/1560-7917.ES.2021.26.29.2100670

407 Dagan N, Barda N, Kepten E, *et al.* BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *N Engl J Med* 2021;**384**:1412–23. doi:10.1056/NEJMoa2101765

408 Chemaitelly H, Yassine HM, Benslimane FM, *et al.* mRNA-1273 COVID-19 vaccine effectiveness against the B.1.1.7 and B.1.351 variants and severe COVID-19 disease in Qatar. *Nat Med* 2021;**27**:1614–21. doi:10.1038/s41591-021-01446-y

409 Lopez Bernal J, Andrews N, Gower C, *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N Engl J Med* 2021;**385**:585–94. doi:10.1056/NEJMoa2108891

410 Bedston S, Akbari A, Jarvis CI, *et al.* COVID-19 vaccine uptake, effectiveness, and waning in 82,959 health care workers: A national prospective cohort study in Wales. *Vaccine* 2022;**40**:1180–9. doi:10.1016/j.vaccine.2021.11.061

411 Waxman JG, Makov-Assif M, Reis BY, *et al.* Comparing COVID-19-related hospitalization rates among individuals with infection-induced and vaccine-induced immunity in Israel. *Nat Commun* 2022;**13**:2202. doi:10.1038/s41467-022-29858-5

412 Gazit S, Shlezinger R, Perez G, *et al.* The Incidence of SARS-CoV-2 Reinfection in Persons With Naturally Acquired Immunity With and Without Subsequent Receipt of a Single Dose of BNT162b2 Vaccine: A Retrospective Cohort Study. *Ann Intern Med* 2022;**175**:674–81. doi:10.7326/M21-4130

413 Feikin DR, Higdon MM, Abu-Raddad LJ, *et al.* Duration of effectiveness of vaccines against SARS-CoV-2 infection and COVID-19 disease: results of a systematic review and meta-regression. *The Lancet* 2022;**399**:924–44. doi:10.1016/S0140-6736(22)00152-0

414 Tartof SY, Slezak JM, Puzniak L, *et al.* Effectiveness of a third dose of BNT162b2 mRNA COVID-19 vaccine in a large US health system: A retrospective cohort study. *The Lancet Regional Health - Americas* 2022;**9**:100198. doi:10.1016/j.lana.2022.100198

415 Price AM, Olson SM, Newhams MM, *et al.* BNT162b2 Protection against the Omicron Variant in Children and Adolescents. *N Engl J Med* 2022;**386**:1899–909. doi:10.1056/NEJMoa2202826

416 Polinski JM, Weckstein AR, Batech M, *et al.* Effectiveness of the Single-Dose Ad26.COV2.S COVID Vaccine. Infectious Diseases (except HIV/AIDS) 2021. doi:10.1101/2021.09.10.21263385

417 Ioannidis JPA. Factors influencing estimated effectiveness of COVID-19 vaccines in non-randomised studies. *BMJ EBM* 2022;:bmjebm-2021-111901. doi:10.1136/bmjebm-2021-111901

418 Fell DB, Dimitris MC, Hutcheon JA, *et al.* Guidance for design and analysis of observational studies of fetal and newborn outcomes following COVID-19 vaccination during pregnancy. *Vaccine* 2021;**39**:1882–6. doi:10.1016/j.vaccine.2021.02.070

419 Dean NE, Hogan JW, Schnitzer ME. Covid-19 Vaccine Effectiveness and the Test-Negative Design. *N Engl J Med* 2021;**385**:1431–3. doi:10.1056/NEJMe2113151

420 Skowronski D. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine: a letter to the Editor. *N Engl J Med* 2021;**384**:1576-. doi:10.1056/NEJMc2036242

421 Tabak YP, Sun X, Brennan TA, *et al.* Incidence and Estimated Vaccine Effectiveness Against Symptomatic SARS-CoV-2 Infection Among Persons Tested in US Retail Locations, May 1 to August 7, 2021. *JAMA Netw Open* 2021;**4**:e2143346. doi:10.1001/jamanetworkopen.2021.43346

422 Barda N, Dagan N, Cohen C, *et al.* Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. *The Lancet* 2021;**398**:2093–100. doi:10.1016/S0140-6736(21)02249-2

423 Hall VJ, Foulkes S, Saei A, *et al.* COVID-19 vaccine coverage in health-care workers in England and effectiveness of BNT162b2 mRNA vaccine against infection (SIREN): a prospective, multicentre, cohort study. *The Lancet* 2021;**397**:1725–35. doi:10.1016/S0140-6736(21)00790-X

424 Pilishvili T, Gierke R, Fleming-Dutra KE, *et al.* Effectiveness of mRNA Covid-19 Vaccine among U.S. Health Care Personnel. *N Engl J Med* 2021;**385**:e90. doi:10.1056/NEJMoa2106599

425 Ostropolets A. Vaccine effectiveness and bias in COVID-19 vaccine effectiveness research. 2022.https://github.com/aostropolets/VaccineEffectiveness (accessed 8 Nov 2022).

426 The Knowledge Base workgroup of the Observational Health Data Sciences and Informatics (OHDSI) collaborative. Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. *J Biomed Semant* 2017;**8**:11. doi:10.1186/s13326-017-0115-3

427 Glasziou P, McCaffery K, Cvejic E, *et al.* Testing behaviour may bias observational studies of vaccine effectiveness. Infectious Diseases (except HIV/AIDS) 2022. doi:10.1101/2022.01.17.22269450

428 Lewnard JA, Tedijanto C, Cowling BJ, *et al.* Measurement of Vaccine Direct Effects Under the Test-Negative Design. *American Journal of Epidemiology* 2018;**187**:2686–97. doi:10.1093/aje/kwy163

429 Hitchings MDT, Lewnard JA, Dean NE, *et al.* Use of recently vaccinated individuals to detect bias in test-negative case–control studies of COVID-19 vaccine effectiveness. *Epidemiology* 2022;**Publish Ahead of Print**. doi:10.1097/EDE.0000000000001484

430 Hadi YB, Thakkar S, Shah-Khan SM, *et al.* COVID-19 Vaccination Is Safe and Effective in Patients With Inflammatory Bowel Disease: Analysis of a Large Multi-institutional Research Network in the United States. *Gastroenterology* 2021;**161**:1336-1339.e3. doi:10.1053/j.gastro.2021.06.014

431 Nunes B, Rodrigues AP, Kislaya I, *et al.* mRNA vaccine effectiveness against COVID-19-related hospitalisations and deaths in older adults: a cohort study based on data linkage of national health registries in Portugal, February to August 2021. *Eurosurveillance* 2021;**26**. doi:10.2807/1560-7917.ES.2021.26.38.2100833

432 Wright BJ, Tideman S, Diaz GA, *et al.* Comparative vaccine effectiveness against severe COVID-19 over time in US hospital administrative data: a case-control study. *The Lancet Respiratory Medicine* 2022;:S221326002200042X. doi:10.1016/S2213-2600(22)00042-X

433 Bodilsen J, Leth S, Nielsen SL, *et al.* Positive Predictive Value of ICD-10 Diagnosis Codes for COVID-19. *CLEP* 2021;**Volume 13**:367–72. doi:10.2147/CLEP.S309840

434 Lynch KE, Viernes B, Gatsby E, *et al.* Positive Predictive Value of COVID-19 ICD-10 Diagnosis Codes Across Calendar Time and Clinical Setting. *CLEP* 2021;**Volume 13**:1011–8. doi:10.2147/CLEP.S335621

435 Ostropolets A, Ryan PB, Schuemie MJ, *et al.* Differential anchoring effects of vaccination comparator selection: characterizing a potential bias due to healthcare utilization in COVID-19 versus influenza. Epidemiology 2021. doi:10.1101/2021.10.07.21264711

436 Gisladottir U, Tatonetti N. Adjusting for Healthcare Utilization Improves the Performance of Self-Controlled Case Series Studies using Electronic Health Records. 2022.

437 Tricoci P. Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines. *JAMA* 2009;**301**:831. doi:10.1001/jama.2009.205

438 Ely JW, Osheroff JA, Ebell MH, *et al.* Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ* 2002;**324**:710. doi:10.1136/bmj.324.7339.710

439 Ely JW, Osheroff JA, Maviglia SM, *et al.* Patient-care questions that physicians are unable to answer. *J Am Med Inform Assoc* 2007;**14**:407–14. doi:10.1197/jamia.M2398

440 Westbrook JI. Do Clinicians Use Online Evidence to Support Patient Care? A Study of 55,000 Clinicians. *Journal of the American Medical Informatics Association* 2003;**11**:113–20. doi:10.1197/jamia.M1385

441 Ru B, Wang X, Yao L. Evaluation of the informatician perspective: determining types of research papers preferred by clinicians. *BMC Medical Informatics and Decision Making* 2017;**17**. doi:10.1186/s12911-017-0463-z

442 Oswald N, Bateman H. Treating individuals according to evidence: why do primary care practitioners do what they do? *J Eval Clin Pract* 2000;**6**:139–48.

443 Paul G. Information Needs in Primary Care: A Survey of Rural and Nonrural Primary Care Physicians. *Studies in Health Technology and Informatics* 2001;:338–42. doi:10.3233/978-1-60750-928-8-338

444 Magrabi F, Westbrook JI, Kidd MR, *et al.* Long-Term Patterns of Online Evidence Retrieval Use in General Practice: A 12-Month Study. *Journal of Medical Internet Research* 2008;**10**:e6. doi:10.2196/jmir.974

445 Clarke MA, Belden JL, Koopman RJ, *et al.* Information needs and information-seeking behaviour analysis of primary care physicians and nurses: a literature review. *Health Information & Libraries Journal* 2013;**30**:178–90. doi:10.1111/hir.12036

446 Bennett NL, Casebeer LL, Kristofco R, *et al.* Family physicians' information seeking behaviors: A survey comparison with other specialties. *BMC Medical Informatics and Decision Making* 2005;**5**. doi:10.1186/1472-6947-5-9

447 Osborn R, Moulds D, Schneider EC, *et al.* Primary Care Physicians In Ten Countries Report Challenges Caring For Patients With Complex Health Needs. *Health Affairs* 2015;**34**:2104–12. doi:10.1377/hlthaff.2015.1018

448 Green ML, Ruff TR. Why do residents fail to answer their clinical questions? A qualitative study of barriers to practicing evidence-based medicine. *Acad Med* 2005;**80**:176–82.

449 Ramos K, Linscheid R, Schafer S. Real-time information-seeking behavior of residency physicians. *Fam Med* 2003;**35**:257–60.

450 Lander B, Balka E. Exploring How Evidence is Used in Care Through an Organizational Ethnography of Two Teaching Hospitals. *Journal of Medical Internet Research* 2019;**21**:e10769. doi:10.2196/10769

451 Tonelli M, Wiebe N, Manns BJ, *et al.* Comparison of the Complexity of Patients Seen by Different Medical Subspecialists in a Universal Health Care System. *JAMA Network Open* 2018;**1**:e184852–e184852. doi:10.1001/jamanetworkopen.2018.4852

452 Aakre CA, Maggio LA, Fiol GD, *et al.* Barriers and facilitators to clinical information seeking: a systematic review. *J Am Med Inform Assoc* 2019;**26**:1129–40. doi:10.1093/jamia/ocz065

453 Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology* 2006;**3**:77–101. doi:10.1191/1478088706qp063oa

454 Ryan GW, Bernard HR. Techniques to Identify Themes. *Field Methods* 2003;**15**:85–109. doi:10.1177/1525822X02239569

455 Caldwell E, Whitehead M, Fleming J, *et al.* Evidence-based practice in everyday clinical practice: Strategies for change in a tertiary occupational therapy department. *Australian Occupational Therapy Journal* 2008;**55**:79–84. doi:10.1111/j.1440-1630.2007.00669.x

456 Jackson R, Feder G. Guidelines for clinical guidelines. *BMJ* 1998;**317**:427–8. doi:10.1136/bmj.317.7156.427

457 Oude Rengerink K, Thangaratinam S, Barnfield G, *et al.* How can we teach EBM in clinical practice? An analysis of barriers to implementation of on-the-job EBM teaching and learning. *Medical Teacher* 2011;**33**:e125–30. doi:10.3109/0142159X.2011.542520

458 Cook DA, Sorensen KJ, Hersh W, *et al.* Features of Effective Medical Knowledge Resources to Support Point of Care Learning: A Focus Group Study. *PLoS ONE* 2013;**8**:e80318. doi:10.1371/journal.pone.0080318

459 Bradley P, Humphris G. Assessing the ability of medical students to apply evidence in practice: the potential of the OSCE. *Med Educ* 1999;**33**:815–7.

460 Benner P. From novice to expert. *Menlo Park* 1984.

461 Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med* 1985;**103**:596–9.

462 Schumock GT, Li EC, Suda KJ, *et al.* National trends in prescription drug expenditures and projections for 2016. *American Journal of Health-System Pharmacy* 2016;**73**:1058–75. doi:10.2146/ajhp160205

463 Stewart WF, Shah NR, Selna MJ, *et al.* Bridging The Inferential Gap: The Electronic Health Record And Clinical Evidence. *Health Affairs* 2007;**26**:w181–91. doi:10.1377/hlthaff.26.2.w181

464 Hajjaj F, Salek M, Basra M, *et al.* Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *Journal of the Royal Society of Medicine* 2010;**103**:178–87. doi:10.1258/jrsm.2010.100104

465 Maggio LA, Aakre CA, Del Fiol G, *et al.* Impact of Clinicians' Use of Electronic Knowledge Resources on Clinical and Learning Outcomes: Systematic Review and Meta-Analysis. *J Med Internet Res* 2019;**21**:e13315. doi:10.2196/13315

466 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 2005;**8**:19–32. doi:10.1080/1364557032000119616

467 Song M, Spallek H, Polk D, *et al.* How information systems should support the information needs of general dentists in clinical settings: suggestions from a qualitative study. *BMC Med Inform Decis Mak* 2010;**10**:7. doi:10.1186/1472-6947-10-7

468 Forsythe DE, Buchanan BG, Osheroff JA, *et al.* Expanding the concept of medical information: An observational study of physicians' information needs. *Computers and Biomedical Research* 1992;**25**:181–200. doi:10.1016/0010-4809(92)90020-B

469 Green ML, Ciampi MA, Ellis PJ. Residents' medical information needs in clinic: are they being met? *Am J Med* 2000;**109**:218–23.

470 Bright TJ, Wong A, Dhurjati R, *et al.* Effect of clinical decision-support systems: a systematic review. *Ann Intern Med* 2012;**157**:29–43. doi:10.7326/0003-4819-157-1-201207030-00450

471 Berner ES. *Clinical decision support systems*. Second Edition. Springer 2007.

472 Middleton B, Sittig DF, Wright A. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision. *Yearbook of Medical Informatics* 2016;**25**:S103–16. doi:10.15265/IYS-2016-s034

473 Zhang Z, Gotz D, Perer A. Iterative cohort analysis and exploration. *Information Visualization* 2015;**14**:289–307. doi:10.1177/1473871614526077

474 Rogers J, Spina N, Neese A, *et al.* Composer—Visual Cohort Analysis of Patient Outcomes. *Applied Clinical Informatics* 2019;**10**:278–85. doi:10.1055/s-0039-1687862

475 Happe A, Drezen E. A visual approach of care pathways from the French nationwide SNDS database - from population to individual records: the ePEPS toolbox. *Fundamental & Clinical Pharmacology* 2018;**32**:81–4. doi:10.1111/fcp.12324

476 Plaisant C, Lam S, Lam SJ, *et al.* Searching electronic health records for temporal patterns in patient histories: a case study with microsoft amalga. *AMIA Annu Symp Proc* 2008;:601–5.

477 Fails J, Karlson A, Shahamat L, *et al.* A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories. In: *2006 IEEE Symposium On Visual Analytics And Technology*. Baltimore, MD, USA: : IEEE 2006. 167–74. doi:10.1109/VAST.2006.261421

478 Lam S, Shneiderman B. PatternFinder in Microsoft Amalga : Temporal Query Formulation and Result Visualization in Action. 2008.

479 Song H, Brennan J, Negahban N. PatternFinder 2 . 0 : Usability Test and Redesign of a Patient History Search System. 2005.

480 Ayewah N, Johnson G, Song H. Patternfinder 3.0: Sparse Temporal Data Visual Query Application.

481 Perer A, Gotz D. Data-driven exploration of care plans for patients. In: *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*. Paris, France: : ACM Press 2013. 439. doi:10.1145/2468356.2468434

482 Morrison JJ, Hostetter J, Wang K, *et al.* Data-driven decision support for radiologists: re-using the National Lung Screening Trial dataset for pulmonary nodule management. *J Digit Imaging* 2015;**28**:18–23. doi:10.1007/s10278-014-9720-1

483 Drezen E, Guyet T, Happe A. From medico-administrative databases analysis to care trajectories analytics: an example with the French SNDS. *Fundamental & Clinical Pharmacology* 2018;**32**:78–80. doi:10.1111/fcp.12323

484 Scailteux L-M, Droitcourt C, Balusson F, *et al.* French administrative health care database (SNDS): The value of its enrichment. *Therapies* 2019;**74**:215–23. doi:10.1016/j.therap.2018.09.072

485 Li P, Yates SN, Lovely JK, *et al.* Patient-like-mine: A real time, visual analytics tool for clinical decision support. In: *2015 IEEE International Conference on Big Data (Big Data)*. 2015. 2865–7. doi:10.1109/BigData.2015.7364104

486 Malik S, Du F, Monroe M, *et al.* Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*. Atlanta, Georgia, USA: : ACM Press 2015. 38–49. doi:10.1145/2678025.2701407

487 Bernard J, Sessler D, May T, *et al.* A Visual-Interactive System for Prostate Cancer Cohort Analysis. *IEEE Computer Graphics and Applications* 2015;**35**:44–55. doi:10.1109/MCG.2015.49

488 Bernard J, Sessler D, Kohlhammer J, *et al.* Using Dashboard Networks to Visualize Multiple Patient Histories: A Design Study on Post-Operative Prostate Cancer. *IEEE Transactions on Visualization and Computer Graphics* 2019;**25**:1615–28. doi:10.1109/TVCG.2018.2803829

489 Finlayson SG, Levy M, Reddy S, *et al.* Toward rapid learning in cancer treatment selection: An analytical engine for practice-based clinical data. *J Biomed Inform* 2016;**60**:104–13. doi:10.1016/j.jbi.2016.01.005

490 Mane KK, Bizon C, Schmitt C, *et al.* VisualDecisionLinc: a visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *J Biomed Inform* 2012;**45**:101–6. doi:10.1016/j.jbi.2011.09.003

491 Nan Cao, Gotz D, Sun J, *et al.* DICON: Interactive Visual Analysis of Multidimensional Clusters. *IEEE Transactions on Visualization and Computer Graphics* 2011;**17**:2581–90. doi:10.1109/TVCG.2011.188

492 Xia E, Liu H, Li J, *et al.* Gathering Real World Evidence with Cluster Analysis for Clinical Decision Support. *Stud Health Technol Inform* 2017;**245**:1185–9.

493 Xia E, Wang K, Zhang Y, *et al.* A Data-driven Clinical Decision Support System for Acute Coronary Syndrome Patient Similarity. In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. Xi'an, China: : IEEE 2019. 1–6. doi:10.1109/ICHI.2019.8904614

494 Gallego B, Walter SR, Day RO, *et al.* Bringing cohort studies to the bedside: framework for a 'green button' to support clinical decision-making. *Journal of Comparative Effectiveness Research* 2015;**4**:191–7. doi:10.2217/cer.15.12

495 Schuler A, Callahan A, Jung K, *et al.* Performing an Informatics Consult: Methods and Challenges. *Journal of the American College of Radiology* 2018;**15**:563–8. doi:10.1016/j.jacr.2017.12.023

496 Gombar S, Callahan A, Califf R, *et al.* It is time to learn from patients like mine. *npj Digital Medicine* 2019;**2**. doi:10.1038/s41746-019-0091-3

497 Yu Y, Liu H, Li J, *et al.* Care pathway workbench: evidence harmonization from guideline and data. *Stud Health Technol Inform* 2014;**205**:23–7.

498 Feinberg BA, Lang J, Grzegorczyk J, *et al.* Implementation of Cancer Clinical Care Pathways: A Successful Model of Collaboration Between Payers and Providers. *Journal of Oncology Practice* 2012;**8**:e38s–43s. doi:10.1200/JOP.2012.000564

499 Nabhan C, Mato AR, Feinberg BA. Clinical pathways in chronic lymphocytic leukemia: Challenges and solutions: Pathways for CLL. *American Journal of Hematology* 2017;**92**:5–6. doi:10.1002/ajh.24589

500 Kreys ED, Koeller JM. Documenting the Benefits and Cost Savings of a Large Multistate Cancer Pathway Program From a Payer's Perspective. *Journal of Oncology Practice* 2013;**9**:e241–7. doi:10.1200/JOP.2012.000871

501 Ellis PG. Development and Implementation of Oncology Care Pathways in an Integrated Care Network: The Via Oncology Pathways Experience. *Journal of Oncology Practice* 2013;**9**:171–3. doi:10.1200/JOP.2013.001020

502 Ellis PG, O'Neil BH, Earle MF, *et al.* Clinical Pathways: Management of Quality and Cost in Oncology Networks in the Metastatic Colorectal Cancer Setting. *Journal of Oncology Practice* 2017;**13**:e522–9. doi:10.1200/JOP.2016.019232

503 Hoverman JR, Klein I, Harrison DW, *et al.* Opening the Black Box: The Impact of an Oncology Management Program Consisting of Level I Pathways and an Outbound Nurse Call System. *Journal of Oncology Practice* 2014;**10**:63–7. doi:10.1200/JOP.2013.001210

504 Hoverman JR, Cartwright TH, Patt DA, *et al.* Pathways, Outcomes, and Costs in Colon Cancer: Retrospective Evaluations in Two Distinct Databases. *Journal of Oncology Practice* 2011;**7**:52s–9s. doi:10.1200/JOP.2011.000318

505 Neubauer MA, Hoverman JR, Kolodziej M, *et al.* Cost Effectiveness of Evidence-Based Treatment Guidelines for the Treatment of Non–Small-Cell Lung Cancer in the Community Setting. *Journal of Oncology Practice* 2010;**6**:12–8. doi:10.1200/JOP.091058

506 Hoverman JR, Klein I, Harrison D, *et al.* Impact of an oncology management program using level I pathways, a nurse call system, and advance care planning. *Journal of Clinical Oncology* 2013;**31**:119–119. doi:10.1200/jco.2013.31.31_suppl.119

507 Goldsmith PJ. NCCN's Value Pathways: The Drive for Quality Cancer Care. *Journal of the National Comprehensive Cancer Network* 2013;**11**:119–20. doi:10.6004/jnccn.2013.0018

508 Kolodziej M, Hoverman JR, Garey JS, *et al.* Benchmarks for Value in Cancer Care: An Analysis of a Large Commercial Population. *Journal of Oncology Practice* 2011;**7**:301–6. doi:10.1200/JOP.2011.000394

509 Espirito JL, Arlen A, Garey JS, *et al.* A quality improvement program to increase performance with clinical oncology pathways. *Journal of Clinical Oncology* 2012;**30**:65–65. doi:10.1200/jco.2012.30.34_suppl.65

510 DeMartino JK, Larsen JK. Equity in Cancer Care: Pathways, Protocols, and Guidelines. *Journal of the National Comprehensive Cancer Network* 2012;**10**:S-1-S-9. doi:10.6004/jnccn.2012.0164

511 Kelly RJ, Forde PM, Elnahal SM, *et al.* Patients and Physicians Can Discuss Costs of Cancer Treatment in the Clinic. *Journal of Oncology Practice* 2015;**11**:308–12. doi:10.1200/JOP.2015.003780

512 *Clinical Practice Guidelines We Can Trust*. Washington, D.C.: : National Academies Press 2011. doi:10.17226/13058

513 Simon G, DiNardo CD, Takahashi K, *et al.* Applying Artificial Intelligence to Address the Knowledge Gaps in Cancer Care. *The Oncologist* 2019;**24**:772–82. doi:10.1634/theoncologist.2018-0257

514 Toni F. Argumentation-based clinical decision support system in ROAD2H. Tempe, Arizona, USA: 2018.

515 Delaney BC, Curcin V, Andreasson A, *et al.* Translational Medicine and Patient Safety in Europe: TRANSFoRm—Architecture for the Learning Health System in Europe. *BioMed Research International* 2015;**2015**:1–8. doi:10.1155/2015/961526

516 Cook DA, Sorensen KJ, Linderbaum JA, *et al.* Information needs of generalists and specialists using online best-practice algorithms to answer clinical questions. *J Am Med Inform Assoc* 2017;**24**:754–61. doi:10.1093/jamia/ocx002

517 Henry C, Etain B, Mathieu F, *et al.* A French network of bipolar expert centres: a model to close the gap between evidence-based medicine and routine practice. *J Affect Disord* 2011;**131**:358–63. doi:10.1016/j.jad.2010.11.013

518 Moja L, Kwag KH, Lytras T, *et al.* Effectiveness of Computerized Decision Support Systems Linked to Electronic Health Records: A Systematic Review and Meta-Analysis. *American Journal of Public Health* 2014;**104**:e12–22. doi:10.2105/AJPH.2014.302164

519 Kilsdonk E, Peute LW, Jaspers MWM. Factors influencing implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis. *International Journal of Medical Informatics* 2017;**98**:56–64. doi:10.1016/j.ijmedinf.2016.12.001

520 Shahmoradi L, Safadari R, Jimma W. Knowledge management implementation and the tools utilized in healthcare for evidence-based decision making: a systematic review. *Ethiopian Journal of Health Sciences* 2017;**27**:541. doi:10.4314/ejhs.v27i5.13

521 Tunis SR, Stryer DB, Clancy CM. Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. *JAMA* 2003;**290**. doi:10.1001/jama.290.12.1624

522 Akyürek ÇE, Sawalha R, Ide S. Factors affecting the decision making process in healthcare institutions. *Academy of Strategic Management Journal* 2015;**14**:1–14.

523 Osop HB. A Practice-Based Evidence Approach For Clinical Decision Support. 2018. doi:10.5204/thesis.eprints.123320

524 Plaisant C, Milash B, Rose A, *et al.* LifeLines: visualizing personal histories. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1996. 221–7.

525 Shahar Y, Cheng C. Intelligent visualization and exploration of time-oriented clinical data. In: Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers. IEEE 1999. 15–31.

526 LePendu P, Iyer SV, Fairon C, *et al.* Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of Biomedical Semantics* 2012;**3**. doi:10.1186/2041-1480-3-S1-S5

527 Jung K, LePendu P, Iyer S, *et al.* Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *Journal of the American Medical Informatics Association* Published Online First: 21 October 2014. doi:10.1136/amiajnl-2014-002902

528 Overhage JM, Ryan PB, Reich CG, *et al.* Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* 2012;**19**:54–60. doi:10.1136/amiajnl-2011-000376

529 Wasserman S, Pattison P, Steinley D. Social Networks. In: Balakrishnan N, Colton T, Everitt B, *et al.*, eds. *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: : John Wiley & Sons, Ltd 2014. doi:10.1002/9781118445112.stat06716

530 Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International Journal of Epidemiology* 2018;**47**:2005–14. doi:10.1093/ije/dyy120

531 Schuemie MJ, Ryan PB, DuMouchel W, *et al.* Interpreting observational studies: why empirical calibration is needed to correct *p* □values. *Statist Med* 2014;**33**:209–18. doi:10.1002/sim.5925

532 Vashisht R, Jung K, Schuler A, *et al.* Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative. *JAMA Network Open* 2018;**1**:e181755. doi:10.1001/jamanetworkopen.2018.1755

533 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. *The Lancet* 2019;**394**:1816–26. doi:10.1016/S0140-6736(19)32317-7

534 Duke JD, Ryan PB, Suchard MA, *et al.* Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network. *Epilepsia* 2017;**58**:e101–6. doi:10.1111/epi.13828

535 Burn E, You SC, Sena A, *et al.* Deep phenotyping of 34,128 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America, Europe and Asia: an international network study. Infectious Diseases (except HIV/AIDS) 2020. doi:10.1101/2020.04.22.20074336

536 Huser V, DeFalco FJ, Schuemie M, *et al.* Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2016;**4**:24. doi:10.13063/2327-9214.1239

537 J. Schuemie M, Soledad Cepede M, A. Suchard M, *et al.* How Confident Are We About Observational Findings in Health Care: A Benchmark Study. *Harvard Data Science Review* Published Online First: 31 January 2020. doi:10.1162/99608f92.147cc28e

538 Madigan D, Stang PE, Berlin JA, *et al.* A Systematic Statistical Approach to Evaluating Evidence from Observational Studies. *Annu Rev Stat Appl* 2014;**1**:11–39. doi:10.1146/annurev-statistics-022513-115645

539 Cuschieri S. The STROBE guidelines. *Saudi journal of anaesthesia* 2019;**13**:S31.

540 Croskerry P. Individual variability in clinical decision making and diagnosis. *Diagnosis: Interpreting the Shadows Oxford, UK: CRC Press, Taylor Francis Group* 2017.

541 Dagliati A, Tibollo V, Sacchi L, *et al.* Big Data as a Driver for Clinical Decision Support Systems: A Learning Health Systems Perspective. *Front Digit Humanit* 2018;**5**:8. doi:10.3389/fdigh.2018.00008

542 Johnson A, Cooper GF, Visweswaran S. Patient-Specific Modeling with Personalized Decision Paths. *AMIA Annu Symp Proc* 2020;**2020**:602–11.

543 Giuse NB, Kafantaris SR, Miller MD, *et al.* Clinical medical librarianship: the Vanderbilt experience. *Bull Med Libr Assoc* 1998;**86**:412–6.

544 Longhurst CA, Harrington RA, Shah NH. A 'Green Button' For Using Aggregate Patient Data At The Point Of Care. *Health Affairs* 2014;**33**:1229–35. doi:10.1377/hlthaff.2014.0099

545 Bedford M, Stevens P, Coulton S, et al. Development of risk models for the prediction of new or worsening acute kidney injury on or during hospital admission: a cohort and nested study. Southampton (UK): : NIHR Journals Library 2016. doi:10.3310/hsdr04060

546 White RH, Garcia M, Sadeghi B, *et al.* Evaluation of the predictive value of ICD-9-CM coded administrative data for venous thromboembolism in the United States. *Thrombosis Research* 2010;**126**:61–7. doi:10.1016/j.thromres.2010.03.009

547 Singh S, Fouayzi H, Anzuoni K, *et al.* Diagnostic Algorithms for Cardiovascular Death in Administrative Claims Databases: A Systematic Review. *Drug Safety* 2019;**42**:515–27. doi:10.1007/s40264-018-0754-z

548 Hemingway H, Asselbergs FW, Danesh J, *et al.* Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European Heart Journal* 2018;**39**:1481–95. doi:10.1093/eurheartj/ehx487

549 Falzer PR. Evidence□based medicine's curious path: From clinical epidemiology to patient□centered care through decision analysis. *Journal of Evaluation in Clinical Practice* Published Online First: 26 August 2020. doi:10.1111/jep.13466

550 Parimbelli E, Marini S, Sacchi L, *et al.* Patient similarity for precision medicine: A systematic review. *J Biomed Inform* 2018;**83**:87–96. doi:10.1016/j.jbi.2018.06.001

Appendix for Chapter 3

Appendix 3.1 Deviations observed in teams' GLP1-RA cohort implementation.

Same as in the master cohort	Deviations from the master cohort
Team 1	
2. Inclusion criteria: At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included any GLP1-RA drugs, including combination drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had T2D drug exposure with drug exposure start within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with T2D diagnosis within [-365; 0] days prior to the index date OR those who had 2 outpatient visits with T2D diagnosis within [-365;0] days prior to the index date
8. Exclusion criteria: excluded patients who had insulin exposure	5. Inclusion criteria: included patients who had CLRD drug exposure with drug exposure start within [-365;0] days prior to the index date OR those who had 1 inpatient visit with

within [-365; -1] days prior to the index date	CLRD diagnosis within [-365; -1] days prior to the index date OR those who had 2 outpatient visits with CLRD diagnosis within [-365;0] days prior to the index date
6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1]	7. Exclusion criteria: excluded patients who had pregnancy codes within [-365;-1] days prior to the index date
	9. Exclusion criteria: excluded patients who had DPP4 exposure within [-all; 0] days prior to the index date
	10. Inclusion criteria: did not add add-on therapy
Team 2	
2. At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included only mono GLP1-RA drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had T2D drug exposure with drug exposure start within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with T2D

	diagnosis within [-365; 0] days prior to the index date OR those who had 2 outpatient visits with T2D diagnosis within [-365;0] days prior to the index date
9. Exclusion criteria: excluded patients who had DPP4 exposure within [-365; 0] days prior to the index date	5. Inclusion criteria: included patients who had CLRD drug exposure with drug exposure start within [-365;0] days prior to the index date OR those who had 1 inpatient visit with CLRD diagnosis within [-365; 0] days prior to the index date OR those who had 2 outpatient visits with CLRD diagnosis within [-365;0] days prior to the index date
6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1]	7. Exclusion criteria: excluded patients who had pregnancy procedure/measurement codes within [-60;60] days or pregnancy diagnosis within [-270;270] days
	8. Exclusion criteria: excluded patients who had insulin exposure within [-all; -1] days prior to the index date
	10. Inclusion criteria: included patients with metformin exposure with exposure start [-all, -1], exposure end [30, all], duration - at least 30 days
Team 3	

1. Inclusion criteria: identified the first GLP1-RA exposure within 2006-2016. Included only mono GLP1-RA drugs	4. Inclusion criteria: included patients who had 1 inpatient visit with T2D drug exposure with drug exposure start within [-365;0] days prior to the index date OR those who had 1 inpatient visit with T2D diagnosis within [-365; 0] days prior to the index date OR those who had 2 outpatient visits with T2D diagnosis within [-365;0] days prior to the index date OR those who had 2 outpatient visits with T2D drug within [-365;0] days prior to the index date
2. At least 365 days of prior observation	5. Inclusion criteria: included patients who had 1 inpatient visit with CLRD drug exposure with drug exposure start within [-365;0] days prior to the index date OR those who had 1 inpatient visit with CLRD diagnosis within [-365; 0] days prior to the index date OR those who had 2 outpatient visits with CRLD diagnosis within [-365;0] days prior to the index date OR those who had 2 outpatient visits with CLRD drug within [-365;0] days prior to the index date
3. Inclusion criteria: Age > 17 at the index date	7. Exclusion criteria: excluded patients who had pregnancy codes within [-365;-1] days prior to the index date

6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1]	8. Exclusion criteria: did not exclude patients with prior insulin exposure
	9. Exclusion criteria: excluded patients who had DPP4 exposure within [-all; 0] days prior to the index date
	10. Inclusion criteria: included patients with metformin exposure with exposure start [-365, -1], exposure end - any, duration - any
Team 4	
2. At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included any GLP1-RA drugs, including combination drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had 1 inpatient visit with T2D drug exposure with drug exposure start within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with T2D diagnosis within [-365; -1] days prior to the index date OR those

	who had 2 T2D diagnosis within [-365;-1] days prior to the index date OR those who had 2 T2D drug exposures within [-365;-1] days prior to the index date
10. Inclusion criteria: included patients with at least 2 antidiabetic drugs with exposure start [-all, 0] and exposure end [0, all]	5. Inclusion criteria: included patients who had 1 inpatient visit with CLRD drug exposure with drug exposure start within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with CLRD diagnosis within [-365; -1] days prior to the index date OR those who had 2 CLRD diagnosis within [-365;-1] days prior to the index date OR those who had 2 CLRD drug exposures within [-365;-1] days prior to the index date
	6. Exclusion criteria: excluded patients with exclusion conditions within [-all;0] days prior to the index date (has to be [-365;-1]. For conditions requiring corticosteroid treatment, required condition and corticosteroid exposure within [-365;-1] day
	7. Exclusion criteria: excluded patients who had pregnancy codes within [-280;280] days prior to the index date (has to be [-180;180])
	8. Exclusion criteria: excluded patients who had insulin exposure within [-all; 0] days prior to the index date
	9. Did not exclude patients with prior DPP4 exposure

Team 5	
2. At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included any GLP1-RA drugs, including combination drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had 1 inpatient visit with T2D diagnosis within [-365; -1] days prior to the index date OR those who had 2 T2D diagnosis within [-365;-1] days prior to the index date OR those who had 1 T2D drug exposure within [-365;-1] days prior to the index date
9. Exclusion criteria: excluded patients who had DPP4 exposure within [-365; 0] days prior to the index date	5. Inclusion criteria: included patients who had 1 inpatient visit with CLRD diagnosis within [-365; -1] days prior to the index date OR those who had 2 CLRD diagnosis within [-365;-1] days prior to the index date OR those who had 1 CLRD drug exposure within [-365;-1] days prior to the index date
8. Exclusion criteria: excluded patients who had insulin exposure	7. Exclusion criteria: excluded patients with pregnancy-related conditions, procedures or measurements at day 0

within [-365; -1] days prior to the index date	
6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1] days prior to the index date	10. Inclusion criteria: included patients who had antidiabetic drug exposure with exposure start [-365, -1], exposure end - any, duration - any
Team 6	
2. At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included any GLP1-RA drugs, including combination drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had 1 inpatient visit with (T2D diagnosis or T2D drug exposure) AND (CLRD diagnosis or CLRD drug exposure) within [-365; -1] days prior to the index date OR those who had 2 outpatient visits with (T2D diagnosis or T2D drug

	exposure) AND (CLRD diagnosis or CLRD drug exposure) within [-365; -1] days prior to the index date
8. Exclusion criteria: excluded patients who had insulin exposure within [-365; -1] days prior to the index date	7. Exclusion criteria: excluded patients with pregnancy-related conditions, procedures or measurements within [-365;-1] days
6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1] days prior to the index date	9. Exclusion criteria: excluded patients who had DPP4 exposure within [-all; 0] days prior to the index date
	10. Inclusion criteria: did not add add-on therapy
Team 7	
2. At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included any GLP1-RA

	drugs, including combination drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had T2D drug exposure with drug exposure start within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with T2D diagnosis within [-365; -1] days prior to the index date OR those who had 2 outpatient visits with T2D diagnosis within [-365;-1] days prior to the index date
8. Exclusion criteria: excluded patients who had insulin exposure within [-365; -1] days prior to the index date	5. Inclusion criteria: included patients who had CLRD drug exposure with drug exposure start within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with CLRD diagnosis within [all; -1] days prior to the index date OR those who had 2 outpatient visits with CLRD diagnosis within [-365;-1] days prior to the index date
6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1] days prior to the index date	7. Exclusion criteria: excluded patients who had pregnancy codes within [-365;0] days prior to the index date
	9. Did not exclude patients with prior DPP4 exposure

	10. Inclusion criteria: did not add add-on therapy
Team 8	
1. Inclusion criteria: identified the first GLP1-RA exposure within 2006-2016. Included only mono GLP1-RA drugs	4. Exclusion criteria: did not exclude prior GLP1-RA exposure within 365 days prior
2. At least 365 days of prior observation	5. Inclusion criteria: included patients who had a CLRD diagnosis within [-365;0] days prior to the index date
3. Inclusion criteria: Age > 17 at the index date	7. Exclusion criteria: excluded patients who had pregnancy codes within [-30;0] days prior to the index date
4. Inclusion criteria: included patients who had a T2D diagnosis within [-365;0] days prior to the index date	8. Exclusion criteria: excluded patients who had insulin exposure within [-365; 0] days prior to the index date

6. Exclusion criteria: excluded patients with exclusion conditions within [-365;0] days prior to the index date	10. Inclusion criteria: did not add add-on therapy
9. Exclusion criteria: excluded patients who had DPP4 exposure within [-365; 0] days prior to the index date	
Team 9	
2. At least 365 days of prior observation	1. Inclusion criteria: identified the first GLP1-RA exposure in the patient history and then excluded those who had the first exposure outside of 2006-2016. Included any GLP1-RA drugs, including combination drugs. Exclusion criteria: excluded patients who had GLP1-RA exposure any time prior [-all, -1]
3. Inclusion criteria: Age > 17 at the index date	4. Inclusion criteria: included patients who had 2 outpatient visits with (T2D drug exposure or T2D diagnosis) within [-365;-1] days prior to the index date OR those who had 1 inpatient

	visit with (T2D drug exposure or T2D diagnosis) within [-365; 1] days prior to the index date
6. Exclusion criteria: excluded patients with exclusion conditions within [-365;-1] days prior to the index date	5. Inclusion criteria: included patients who had 2 outpatient visits with (CLRD drug exposure or CLRD diagnosis) within [-365;-1] days prior to the index date OR those who had 1 inpatient visit with (CLRD drug exposure or CLRD diagnosis) within [-365; 1] days prior to the index date
8. Exclusion criteria: excluded patients who had insulin exposure within [-365; -1] days prior to the index date	7. Exclusion criteria: excluded patients who had pregnancy codes within [-365;-1] days prior to the index date
9. Exclusion criteria: excluded patients who had DPP4 exposure within [-365; 0] days prior to the index date	10. Inclusion criteria: included patients who another antidiabetic drug exposure with exposure start [-all, 1], exposure end - any, duration - any

Appendix 3.2 Patient characteristics in the master implementation and teams' implementations.

	Master	**Team 1**	**Team 2**	**Team 3**	**Team 4**	**Team 5**	**Team 6**	**Team 7**	**Team 8**	**Team 9**
Total, n	6356	43641	18701	11260	8417	3126	2159	63619	12618	31253
Female patients, n (%)	4207 (66.2%)	28396 (65.1%)	11575 (61.9%)	7594 (67.4%)	4903 (58.3%)	1794 (57.4%)	1460 (67.6%)	39740 (62.5%)	8531 (67.6%)	19860 (63.5%)
Age groups, n (%)										
15 - 19	9 (0.1%)	98 (0.2%)	32 (0.2%)	20 (0.2%)	<5 (<0.1%)	<5 (<0.1%)	<5 (<0.1%)	104 (0.2%)	14 (0.1%)	46 (0.2%)
20 - 24	35 (0.6%)	309 (0.7%)	96 (0.5%)	74 (0.7%)	12 (0.1%)	20 (0.6%)	8 (0.4%)	329 (0.5%)	78 (0.6%)	123 (0.4%)
25 - 29	39 (0.6%)	543 (1.2%)	185 (1.0%)	127 (1.1%)	56 (0.7%)	33 (1.1%)	12 (0.6%)	645 (1.0%)	120 (1.0%)	222 (0.7%)
30 - 34	121 (1.9%)	1,398 (3.2%)	512 (2.7%)	338 (3.0%)	158 (1.9%)	86 (2.8%)	36 (1.7%)	1,731 (2.7%)	308 (2.4%)	695 (2.2%)
35 - 39	282 (4.4%)	2,863 (6.6%)	1,086 (5.8%)	726 (6.5%)	389 (4.6%)	140 (4.5%)	117 (5.4%)	3,710 (5.8%)	647 (5.1%)	1,674 (5.4%)

40 - 44	608 (9.6%)	4,685 (10.7%)	1,908 (10.2%)	1,255 (11.2%)	750 (8.9%)	253 (8.1%)	200 (9.3%)	6,374 (10.0%)	1,238 (9.8%)	3,073 (9.8%)
45 - 49	897 (14.1%)	6,778 (15.5%)	2,905 (15.5%)	1,758 (15.6%)	1,222 (14.5%)	379 (12.1%)	303 (14.0%)	9,634 (15.1%)	1,865 (14.8%)	4,800 (15.4%)
50 - 54	1,322 (20.8%)	8,736 (20.0%)	3,907 (20.9%)	2,304 (20.5%)	1,744 (20.7%)	543 (17.4%)	444 (20.6%)	12,727 (20.0%)	2,597 (20.6%)	6,698 (21.4%)
55 - 59	1,500 (23.6%)	9,343 (21.4%)	4,378 (23.4%)	2,424 (21.5%)	2,115 (25.1%)	780 (25.0%)	528 (24.5%)	14,228 (22.4%)	2,850 (22.6%)	7,492 (24.0%)
60 - 64	1,448 (22.8%)	8,291 (19.0%)	3,540 (18.9%)	2,071 (18.4%)	1,877 (22.3%)	820 (26.2%)	476 (22.1%)	13,143 (20.7%)	2,660 (21.1%)	6,101 (19.5%)
65 - 69	95 (1.5%)	597 (1.4%)	152 (0.8%)	163 (1.5%)	90 (1.1%)	68 (2.2%)	31 (1.4%)	994 (1.6%)	241 (1.9%)	329 (1.1%)
Patients with a condition record within a year prior to the index date, n (%)										
Asthma	4,059 (63.9%)	6,463 (14.8%)	2,596 (13.9%)	3,391 (30.1%)	1,918 (22.8%)	491 (15.7%)	1,290 (59.8%)	9,199 (14.5%)	7,839 (62.1%)	6,842 (21.9%)

Chronic obstructive lung disease	1,452 (22.8%)	2,132 (4.9%)	808 (4.3%)	1,203 (10.7%)	780 (9.3%)	286 (9.2%)	654 (30.3%)	3,042 (4.8%)	2,603 (20.6%)	2,477 (7.9%)
Hypertensive disorder	4,497 (70.8%)	27,373 (62.7%)	11,694 (62.5%)	7,723 (68.6%)	6,172 (73.3%)	2,095 (67%)	1670 (77.3%)	41,603 (65.4%)	8,821 (69.9%)	20,911 (66.9%)
Congestive heart failure	274 (4.3%)	823 (1.9%)	322 (1.7%)	312 (2.8%)	371 (4.4%)	192 (6.1%)	146 (6.8%)	1,190 (1.9%)	444 (3.5%)	802 (2.6%)
Coronary arteriosclerosis	723 (11.4%)	3,413 (7.8%)	1,438 (7.7%)	1,043 (9.3%)	1,075 (12.8%)	526 (16.8%)	301 (13.9%)	5,347 (8.4%)	1,254 (9.9%)	2,916 (9.3%)
Acute myocardial infarction	69 (1.1%)	264 (0.6%)	116 (0.6%)	107 (0.9%)	102 (1.2%)	47 (1.5%)	38 (1.8%)	413 (0.7%)	107 (0.9%)	240 (0.8%)
Cardiac arrhythmia	484 (7.6%)	2,258 (5.2%)	907 (4.9%)	774 (6.9%)	649 (7.7%)	275 (8.8%)	217 (10.1%)	3,504 (5.5%)	969 (7.7%)	1,963 (6.3%)
Hyperlipidemia	4,183 (65.8%)	26,436 (60.6%)	11,464 (61.3%)	7,381 (65.6%)	5,905 (70.2%)	1,984 (63.5%)	1,509 (69.9%)	40,811 (64.1%)	8,474 (67.2%)	20,190 (64.6%)

Obstructive sleep apnea syndrome	1,494 (23.5%)	7,533 (17.3%)	3,137 (16.8%)	2,442 (21.7%)	1,800 (21.4%)	504 (16.1%)	578 (26.8%)	11,032 (17.3%)	2,930 (23.2%)	5,802 (18.6%)
Cerebral infarction	54 (0.9%)	255 (0.6%)	102 (0.5%)	99 (0.9%)	121 (1.4%)	50 (1.6%)	32 (1.5%)	390 (0.6%)	93 (0.7%)	226 (0.7%)
Hypothyroidism	1,115 (17.5%)	7,468 (17.1%)	2,903 (15.5%)	2,151 (19.1%)	1,263 (15%)	496 (15.9%)	425 (19.7%)	10,721 (16.9%)	2,395 (19%)	5,271 (16.9%)
Pneumonia	471 (7.4%)	1,590 (3.6%)	626 (3.4%)	703 (6.2%)	496 (5.9%)	177 (5.7%)	214 (9.9%)	2,229 (3.5%)	823 (6.5%)	1,324 (4.2%)
Depressive disorder	1,031 (16.2%)	6,064 (13.9%)	2,353 (12.6%)	1,979 (17.6%)	1,280 (15.2%)	466 (14.9%)	438 (20.3%)	8,536 (13.4%)	2,192 (17.4%)	4,488 (14.4%)
Chronic nonalcoholic liver disease	275 (4.3%)	1,512 (3.5%)	622 (3.3%)	508 (4.5%)	380 (4.5%)	131 (4.2%)	118 (5.5%)	2,288 (3.6%)	523 (4.1%)	1,251 (4%)
Obesity	1,967 (31%)	11,130 (25.5%)	4,488 (24%)	3,471 (30.8%)	2,570 (30.5%)	697 (22.3%)	880 (40.8%)	15,,627 (24.6%)	4,028 (31.9%)	7,971 (25.5%)

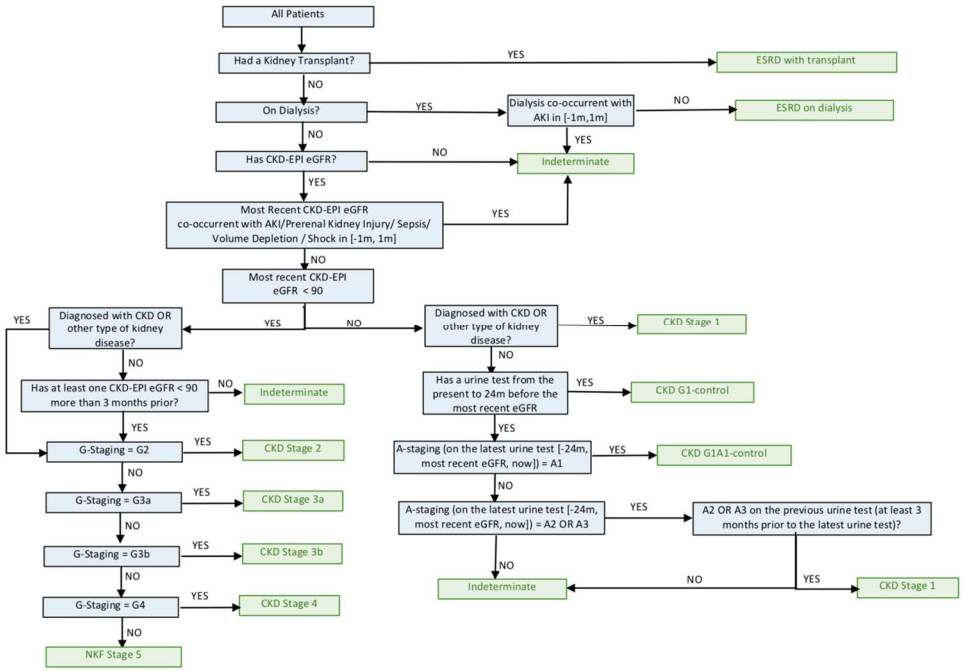
Glaucoma	330 (5.2%)	1,768 (4%)	789 (4.2%)	488 (4.3%)	445 (5.3%)	223 (7.1%)	101 (4.7%)	2,814 (4.4%)	605 (4.8%)	1,416 (4.5%)
Hypoglycemia	51 (0.8%)	604 (1.4%)	198 (1.1%)	157 (1.4%)	94 (1.1%)	73 (2.3%)	28 (1.3%)	745 (1.2%)	133 (1%)	304 (1%)
Patients with a drug exposure record within a year prior to the index date, n (%)										
Metformin	5,858 (92.2%)	37,353 (85.6%)	17,464 (93.4%)	11,260 (100%)	6,,852 (81.4%)	2106 (67.4%)	1,797 (83.2%)	54,742 (86.1%)	9,907 (78.5%)	26,151 (83.7%)
Sulfonylureas	2,888 (45.4%)	15,284 (35%)	8,366 (44.7%)	3,824 (34%)	3,731 (44.3%)	1,069 (34.2%)	880 (40.8%)	25,061 (39.4%)	4,390 (34.8%)	12,980 (41.5%)
Sodium-glucose co-transporter 2 (SGLT2) inhibitors	596 (9.4%)	2,860 (6.6%)	1,620 (8.7%)	896 (8%)	1,052 (12.5%)	141 (4.5%)	207 (9.6%)	6,267 (9.8%)	1,152 (9.1%)	2,746 (8.8%)
Thiazolidinediones	1460 (23%)	9,,506 (21.8%)	5,046 (27%)	1,957 (17.4%)	1,772 (21%)	888 (28.4%)	397 (18.4%)	13,503 (21.2%)	2,027 (16.1%)	7,301 (23.4%)

Appendix 3.3 Unit conversion table for units associated with kidney function measurements.

Measurement	Source unit	Conversion factor	Target unit	Measurement	Source unit	Conversion factor	Target unit
creatinine	ng/ml	0.0001	mg/dl	albumin and protein	mg/g	1	mg/l
creatinine	mcmol/l	0.0113	mg/dl	albumin and protein	mg/24h	1	mg/l
creatinine	mol/l	11300	mg/dl	albumin and protein	mg/dl	0.1	mg/l
creatinine	mmol/l	11.3	mg/dl	albumin and protein	g/l	0.001	mg/l
creatinine	g/l	1000	mg/dl	albumin and protein	mg/ml	0.001	mg/l

heigh	m	100	cm	albumin and protein	mmol/l	0.00067	mg/l
height	inch	2.54	cm				

Appendix 3.4. Flowchart for the gold standard algorithm for chronic kidney disease.



Appendix for Chapter 4

Appendix 4.1 Demographic characteristics of included patient populations stratified by database.

	CCAE	CUMC	MDCD	MDCR	OPTUM EHR	OPTUM SES	IPCI	SIDIAP	DA FRANCE	DA GERMANY	AUSTRALIA EMR	JMDC
Total, n	25,315,777	1,164,196	12,966,011	1,533,709	40,955,085	18,643,608	1,536,283	2,217,536	1,746,371	9,295,525	252,212	6,501,991
Sex, n (%)												
Female	13,037,440 (51.5)	693,190 (59.5)	7,322,471 (56.5)	849,301 (55.4)	23,220,748 (56.7)	9,595,675 (51.5)	783,660 (51.0)	1,120,373 (50.5)	926,180 (53.0)	5,340,273 (57.5)	137,203 (54.4)	2,926,702 (45.0)
Age group, n (%)												
1-5	1,256,501 (5.0)	40,678 (3.5)	1,755,796 (13.5)	0	1,852,425 (4.5)	627,032 (3.4)	78,848 (5.1)	99,838 (4.5)	99,309 (5.7)	308,728 (3.1)	13,430 (5.1)	414,167 (6.4)
6-17	4,122,110 (16.3)	105,520 (9.1)	4,188,247 (32.3)	0	4,773,000 (11.7)	1,930,638 (10.4)	211,037 (13.7)	260,102 (11.7)	268,591 (15.4)	823,235 (8.9)	31,780 (12.6)	1,044,041 (16.1)
18-34	6,395,387 (25.3)	199,020 (17.1)	2,885,991 (22.3)	0	8,182,549 (20.0)	3,331,356 (17.9)	304,971 (19.9)	374,994 (16.9)	328,759 (18.8)	1,411,620 (15.2)	50,995 (20.2)	1,533,866 (23.6)

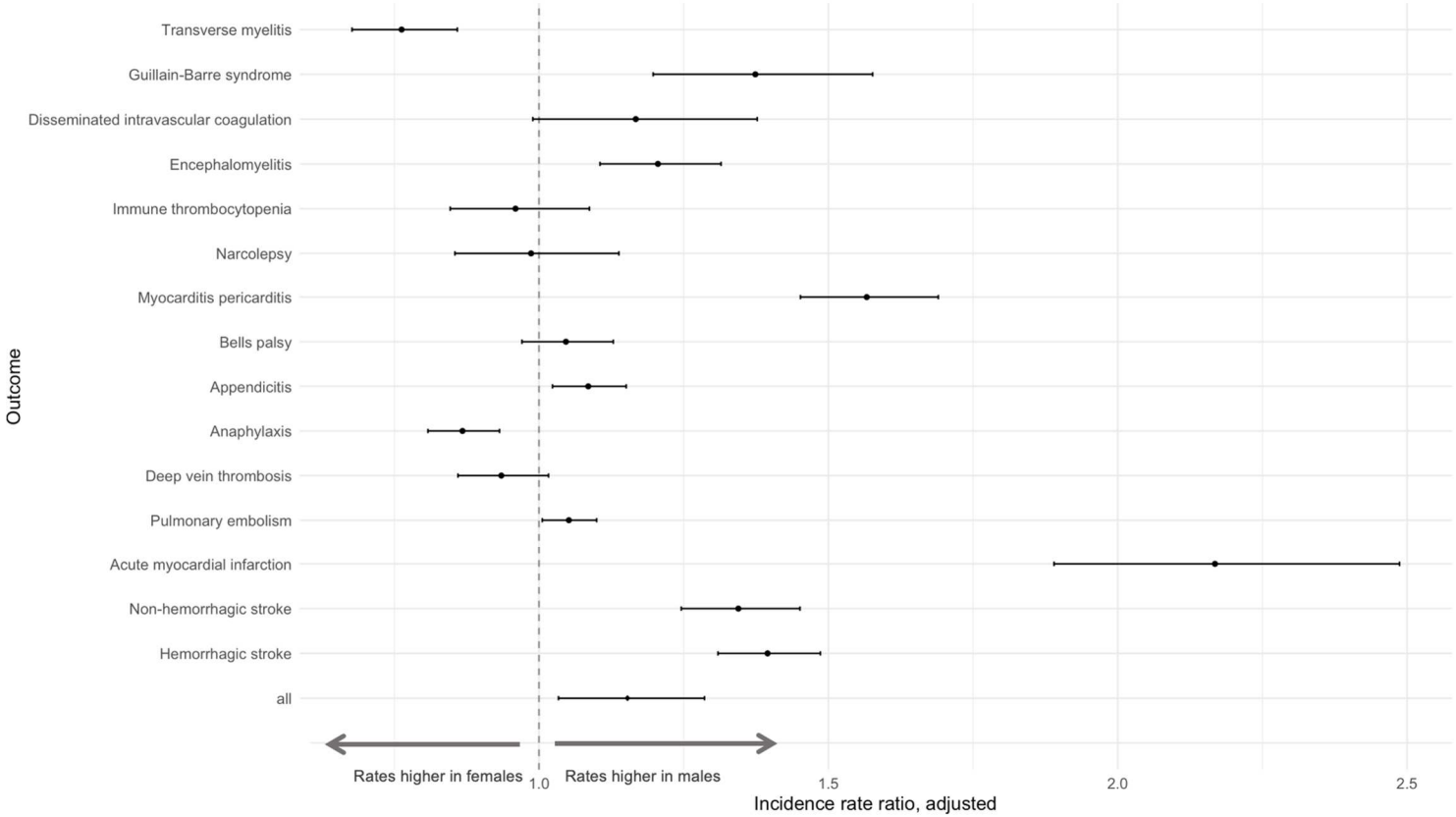
35-54	8,096,864 (32.0)	300,818 (25.8)	2,006,493 (15.5)	0	10,737,664 (26.2)	4,389,220 (23.5)	394,868 (25.7)	663,537 (29.9)	446,804 (25.6)	2,338,535 (25.2)	69,872 (27.7)	2,330,010 (35.8)
55-64	4,716,207 (18.6)	183,612 (15.8)	1,004,957 (7.8)	0	6,655,199 (16.3)	2,384,571 (12.8)	219,990 (14.1)	288,494 (13.0)	229,016 (13.1)	1,580,565 (17.0)	36,329 (14.4)	880,065 (13.5)
65-74	728,708 (2.9)	171,940 (14.8)	633,262 (4.9)	733,157 (47.8)	4,829,968 (11.8)	3,106,611 (16.7)	180,581 (11.8)	246,763 (11.1)	197,816 (11.3)	1,279,048 (13.8)	27,272 (10.8)	279,277 (4.3)
75-84	0	110,883 (9.5)	341,267 (2.6)	536,970 (35.0)	2,652,453 (6.5)	1,985,356 (10.7)	104,288 (6.8)	180,903 (8.2)	117,067 (6.7)	1,191,402 (12.8)	15,319 (6.1)	20,565 (0.3)
≥85	0	51,725 (4.4)	149,998 (1.2)	263,582 (17.2)	1,271,827 (3.1)	888,824 (4.8)	41,700 (2.7)	102,905 (4.6)	59,009 (3.4)	362,392 (3.9)	7,215 (2.9)	0
Race, n (%)												
White		427,525 (38.4)	5,747,120 (48.9)		27,337,580 (73.2)	10,731,895 (62.3)						
Black		105,201 (9.5)	3,948,066 (33.6)		3,925,225 (10.5)	1,576,310 (9.2)						

Appendix 4.2 Pooled age-adjusted incidence rate ratios for race and sex comparison, from meta-analyses, IRR and 95% CI.

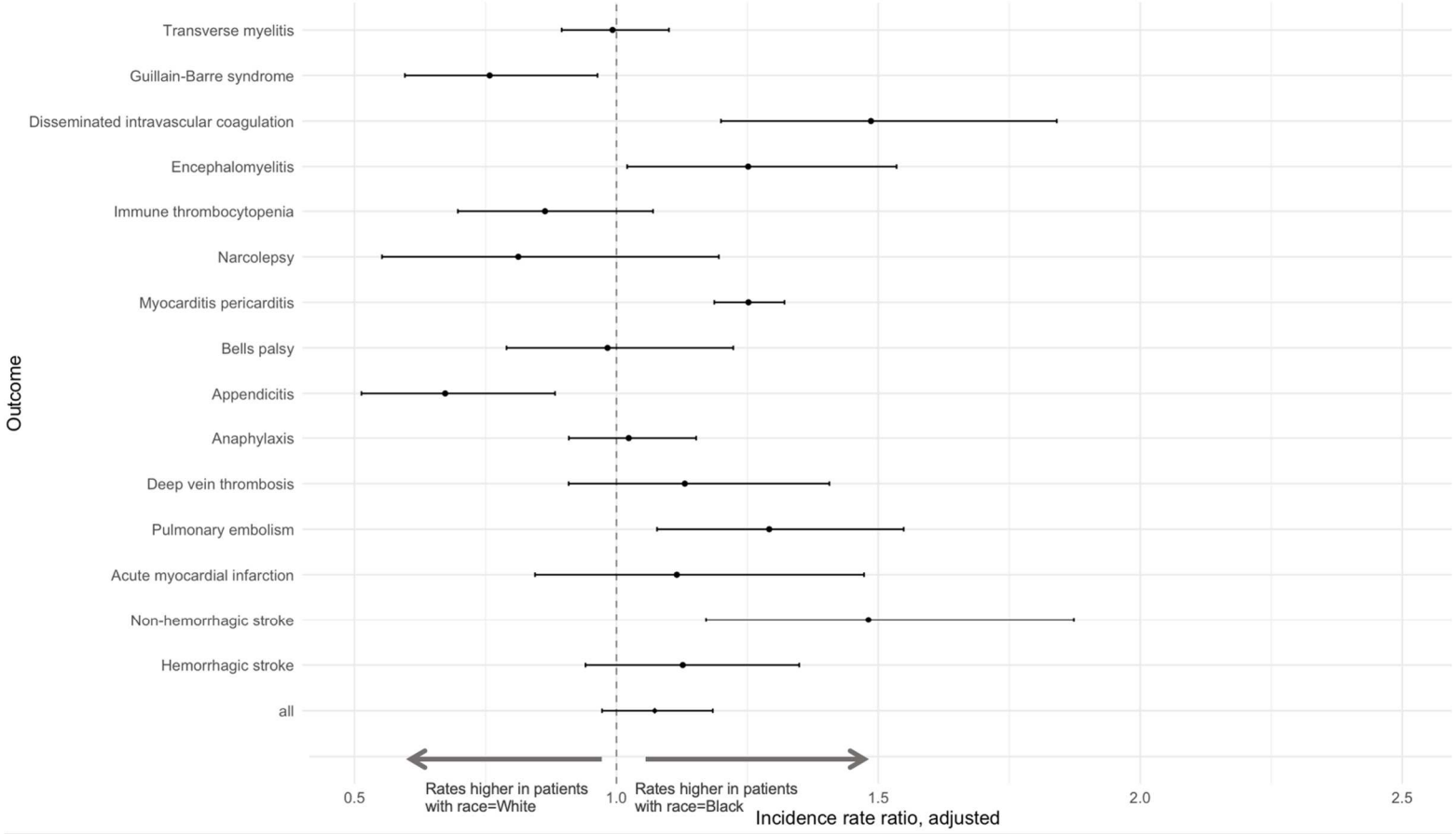
Outcome	Male versus female	Patients with race=Black versus patients with race=White
Acute myocardial infarction	2.17 (1.89-2.49)	1.12 (0.84-1.47)
Anaphylaxis	0.87 (0.81-0.93)	1.02 (0.91-1.15)
Appendicitis	1.09 (1.02-1.15)	0.67 (0.51-0.88)
Bell's palsy	1.05 (0.97-1.13)	0.98 (0.79-1.22)
Deep vein thrombosis	0.93 (0.86-1.02)	1.13 (0.91-1.41)
Disseminated intravascular coagulation	1.17 (0.99-1.38)	1.49 (1.2-1.84)
Encephalomyelitis	1.21 (1.11-1.31)	1.25 (1.02-1.53)
Guillain-Barre syndrome	1.37 (1.2-1.58)	0.76 (0.6-0.96)
Hemorrhagic stroke	1.39 (1.31-1.49)	1.13 (0.94-1.35)
Immune thrombocytopenia	0.96 (0.85-1.09)	0.86 (0.7-1.07)
Myocarditis and pericarditis	1.57 (1.45-1.69)	1.25 (1.19-1.32)
Narcolepsy	0.99 (0.85-1.14)	0.81 (0.55-1.2)
Non-hemorrhagic stroke	1.34 (1.25-1.45)	1.48 (1.17-1.87)

Pulmonary embolism	1.05 (1.01-1.1)	1.29 (1.08-1.55)
Transverse myelitis	0.76 (0.68-0.86)	0.99 (0.9-1.1)
all	1.15 (1.03-1.29)	1.07 (0.97-1.18)

Appendix 4.3 Pooled age-adjusted incidence rate ratios (incidence rates for male versus female patients) from meta-analyses, IRR and 95% CI.



Appendix 4.4 Pooled age-adjusted incidence rate ratios (incidence rates for patients with race=Black versus white patients with race=White) from meta-analyses, IRR and 95% CI.

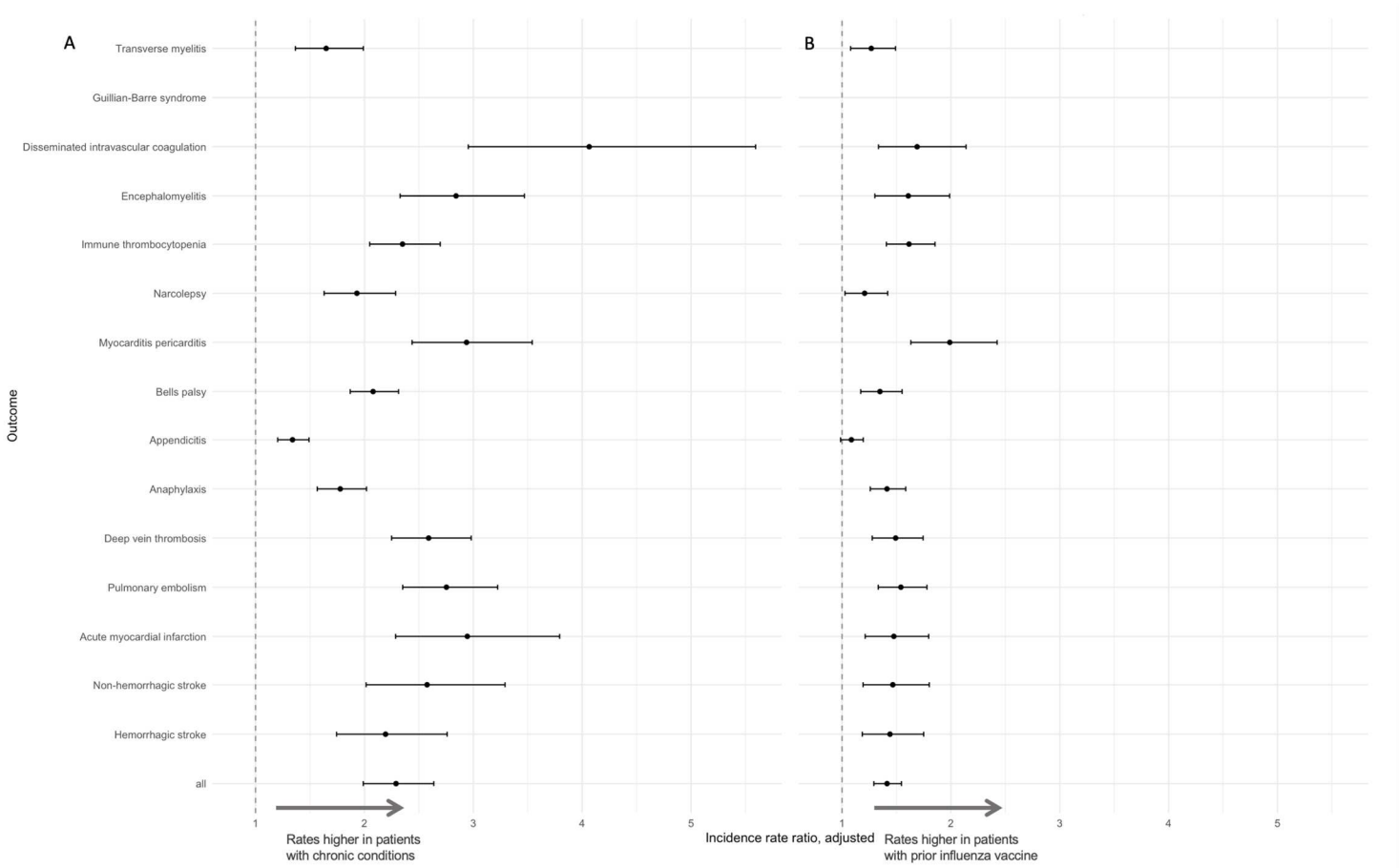


Appendix 4.5. Incidence rate ratios of incidence rates for patient subgroups from meta-analysis, IRR and 95% CI.

Outcome	Patients with prior influenza vaccine versus general population	Patients with chronic conditions versus patients with no chronic conditions
Acute myocardial infarction	1.48 (1.21-1.8)	2.94 (2.29-3.79)
Anaphylaxis	1.41 (1.26-1.59)	1.78 (1.57-2.02)
Appendicitis	1.09 (0.99-1.2)	1.34 (1.2-1.49)
Bell's palsy	1.35 (1.17-1.55)	2.08 (1.87-2.31)
Deep vein thrombosis	1.49 (1.28-1.74)	2.59 (2.25-2.98)
Disseminated intravascular coagulation	1.69 (1.33-2.14)	4.06 (2.95-5.59)
Encephalomyelitis	1.61 (1.3-1.99)	2.84 (2.33-3.47)
Guillain-Barre syndrome	1.01 (0.84-1.22)	2.03 (1.61-2.56)
Hemorrhagic stroke	1.44 (1.19-1.75)	2.19 (1.74-2.76)
Immune thrombocytopenia	1.62 (1.41-1.85)	2.35 (2.05-2.7)
Myocarditis and pericarditis	1.99 (1.63-2.43)	2.94 (2.44-3.54)
Narcolepsy	1.21 (1.03-1.42)	1.93 (1.63-2.29)

Non-hemorrhagic stroke	1.47 (1.19-1.8)	2.58 (2.01-3.29)
Pulmonary embolism	1.54 (1.33-1.78)	2.75 (2.35-3.22)
Transverse myelitis	1.27 (1.08-1.49)	1.65 (1.37-1.99)
All	1.41 (1.29-1.55)	2.29 (1.99-2.64)

Appendix 4.6 Comparison of incidence rates for patients with chronic conditions versus patients with no chronic conditions (A) and patients with prior influenza vaccination versus general population (B) from meta-analysis, IRR and 95% CI.



Appendix 4.7 Pooled age-adjusted incidence rate ratios (incidence rates of outcomes when entering the cohort on a visit versus entering on January 1st in patients with a visit in the next year) from meta-analyses, IRR and 95% CI.

Outcome	Time at risk 0-1 days	Time at risk 1-28 days	Time at risk 1-42 days	Time at risk 1-90 days	Time at risk 1-365 days	Time at risk 0-365 days
Acute myocardial infarction	24.72 (14.35-42.6)	1.23 (1.12-1.35)	1.08 (0.99-1.18)	0.9 (0.82-1)	0.83 (0.76-0.9)	0.98 (0.95-1.01)
Anaphylaxis	49.94 (33.83-73.73)	1.48 (1.24-1.77)	1.37 (1.18-1.58)	1.13 (1.02-1.24)	0.91 (0.88-0.94)	1.09 (1.03-1.16)
Appendicitis	51.79 (42.26-63.47)	1.17 (1.01-1.36)	1.01 (0.9-1.14)	0.83 (0.77-0.9)	0.79 (0.74-0.83)	1.03 (0.99-1.07)
Bell's palsy	34.17 (25.9-45.07)	1.19 (0.95-1.49)	1.05 (0.88-1.27)	0.9 (0.79-1.01)	0.86 (0.83-0.89)	1.03 (0.98-1.09)
Deep vein thrombosis	31.13 (22.48-43.1)	1.42 (1.25-1.61)	1.28 (1.16-1.42)	1.09 (1.02-1.16)	0.91 (0.89-0.93)	1.03 (1-1.07)
Disseminated intravascular coagulation	24.51 (15.31-39.23)	2.04 (1.7-2.45)	1.81 (1.53-2.14)	1.44 (1.26-1.64)	1.02 (0.94-1.09)	1.13 (1.07-1.19)

Encephalomyelitis	16.17 (10.09-25.89)	1.92 (1.72-2.14)	1.59 (1.47-1.71)	1.24 (1.15-1.33)	0.93 (0.88-0.97)	1 (0.98-1.03)
Guillain-Barre syndrome	20.77 (15.14-28.49)	1.71 (1.46-2)	1.43 (1.23-1.65)	1.08 (0.98-1.19)	0.87 (0.82-0.93)	0.99 (0.95-1.02)
Hemorrhagic stroke	27.14 (17.21-42.8)	1.51 (1.38-1.66)	1.3 (1.18-1.42)	1.03 (0.93-1.14)	0.86 (0.78-0.94)	1.01 (0.98-1.04)
Immune thrombocytopenia	25.39 (17.57-36.7)	1.49 (1.26-1.77)	1.33 (1.17-1.52)	1.04 (0.96-1.13)	0.89 (0.86-0.92)	1 (0.97-1.04)
Myocarditis and pericarditis	26.79 (17.04-42.13)	1.47 (1.25-1.73)	1.32 (1.16-1.5)	1.03 (0.93-1.13)	0.9 (0.86-0.93)	1.03 (0.98-1.08)
Narcolepsy	33.25 (22.52-49.07)	1.15 (1.04-1.27)	1.06 (0.98-1.14)	0.96 (0.92-1.01)	0.88 (0.85-0.91)	1 (0.96-1.03)
Non-hemorrhagic stroke	24.01 (13.42-42.94)	1.34 (1.25-1.42)	1.18 (1.12-1.24)	0.96 (0.91-1.02)	0.84 (0.79-0.9)	0.98 (0.96-1.01)
Pulmonary embolism	24.33 (17.3-34.22)	1.41 (1.25-1.59)	1.28 (1.16-1.41)	1.06 (1-1.13)	0.91 (0.89-0.94)	1.03 (1.01-1.06)

Transverse myelitis	17.33 (10-30.02)	1.54 (1.25-1.89)	1.36 (1.17-1.58)	1.03 (0.92-1.16)	0.88 (0.86-0.91)	0.97 (0.94-1.01)
All	28.04 (23.11-34.03)	1.45 (1.33-1.57)	1.28 (1.18-1.38)	1.04 (0.98-1.1)	0.89 (0.87-0.91)	1.02 (1-1.03)

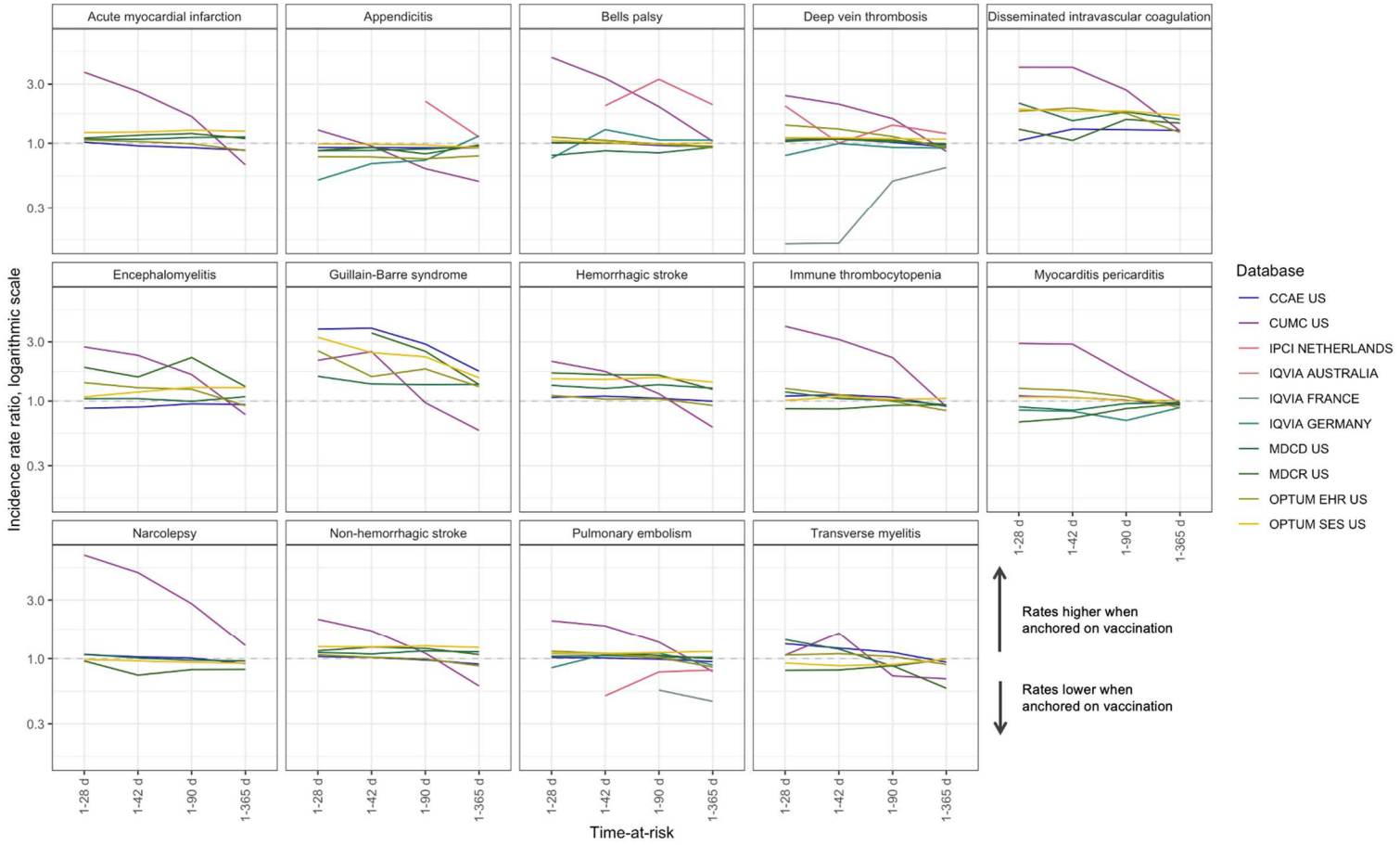
Appendix 4.8 Pooled age-adjusted incidence rate ratios (incidence rates of outcomes when entering the cohort on an influenza vaccination versus on January 1st with an influenza vaccination in the next year, from meta-analyses, IRR and 95% CI.

Outcome	**Time at risk 0-1 day**	**Time at risk 1-28 days**	**Time at risk 1-42 days**	**Time at risk 1-90 days**	**Time at risk 1-365 days**
Acute myocardial infarction	3.91 (1.46-10.46)	1.28 (1.09-1.51)	1.22 (1.06-1.4)	1.15 (1.01-1.31)	0.97 (0.82-1.13)
Anaphylaxis	13.61 (8.97-20.65)	1.12 (1.05-1.2)	1.12 (1.06-1.19)	1.05 (1.01-1.09)	0.94 (0.92-0.96)
Appendicitis	11.65 (6.81-19.94)	0.9 (0.83-0.99)	0.9 (0.83-0.97)	0.85 (0.78-0.94)	0.87 (0.81-0.94)

Bell's palsy	8.84 (5.4-14.49)	1.18 (0.97-1.44)	1.12 (0.97-1.3)	1.01 (0.92-1.11)	0.95 (0.92-0.99)
Deep vein thrombosis	12.19 (8.13-18.29)	1.21 (1.07-1.36)	1.19 (1.09-1.3)	1.09 (1.04-1.15)	0.95 (0.89-1.01)
Disseminated intravascular coagulation	6.36 (3.62-11.18)	1.79 (1.35-2.38)	1.73 (1.36-2.2)	1.71 (1.49-1.96)	1.4 (1.23-1.59)
Encephalomyelitis	2.02 (0.7-5.83)	1.27 (0.96-1.67)	1.22 (0.99-1.51)	1.23 (1.03-1.47)	1.04 (0.89-1.21)
Guillain-Barre syndrome	5.18 (1.85-14.47)	2.82 (2.07-3.86)	2.2 (1.59-3.04)	1.99 (1.54-2.56)	1.4 (1.17-1.66)
Hemorrhagic stroke	4.5 (1.52-13.29)	1.38 (1.16-1.64)	1.32 (1.11-1.57)	1.26 (1.05-1.52)	1.04 (0.86-1.27)
Immune thrombocytopenia	16.21 (7.73-33.99)	1.19 (1-1.43)	1.14 (0.99-1.31)	1.06 (0.96-1.18)	0.92 (0.84-1)
Myocarditis and pericarditis	7.11 (3.54-14.3)	1.13 (0.93-1.38)	1.1 (0.92-1.32)	1.02 (0.94-1.11)	0.94 (0.9-0.99)

Narcolepsy	18.51 (9.64-35.53)	1.05 (0.94-1.16)	1 (0.9-1.11)	0.97 (0.91-1.04)	0.92 (0.9-0.95)
Non-hemorrhagic stroke	4.17 (1.6-10.85)	1.2 (1.07-1.35)	1.17 (1.05-1.3)	1.1 (0.98-1.23)	0.95 (0.81-1.11)
Pulmonary embolism	10.04 (6.56-15.36)	1.11 (1.03-1.2)	1.09 (1.03-1.16)	1.05 (0.99-1.11)	0.93 (0.84-1.04)
Transverse myelitis	8.7 (2.75-27.48)	1.08 (0.86-1.35)	1.05 (0.87-1.25)	0.98 (0.86-1.12)	0.93 (0.87-1)
All	8.68 (6.74-11.17)	1.21 (1.12-1.32)	1.16 (1.08-1.24)	1.11 (1.04-1.18)	0.97 (0.94-1.01)

Appendix 4.9 Incidence rate ratio of incidence rate of outcomes when entering the cohort on an influenza vaccination versus entering on January 1st in patients with an influenza vaccination in the next year, time-at-risk 1-28 days, 1-42 days, 1-90 days and 1-365 days time-at-risk.

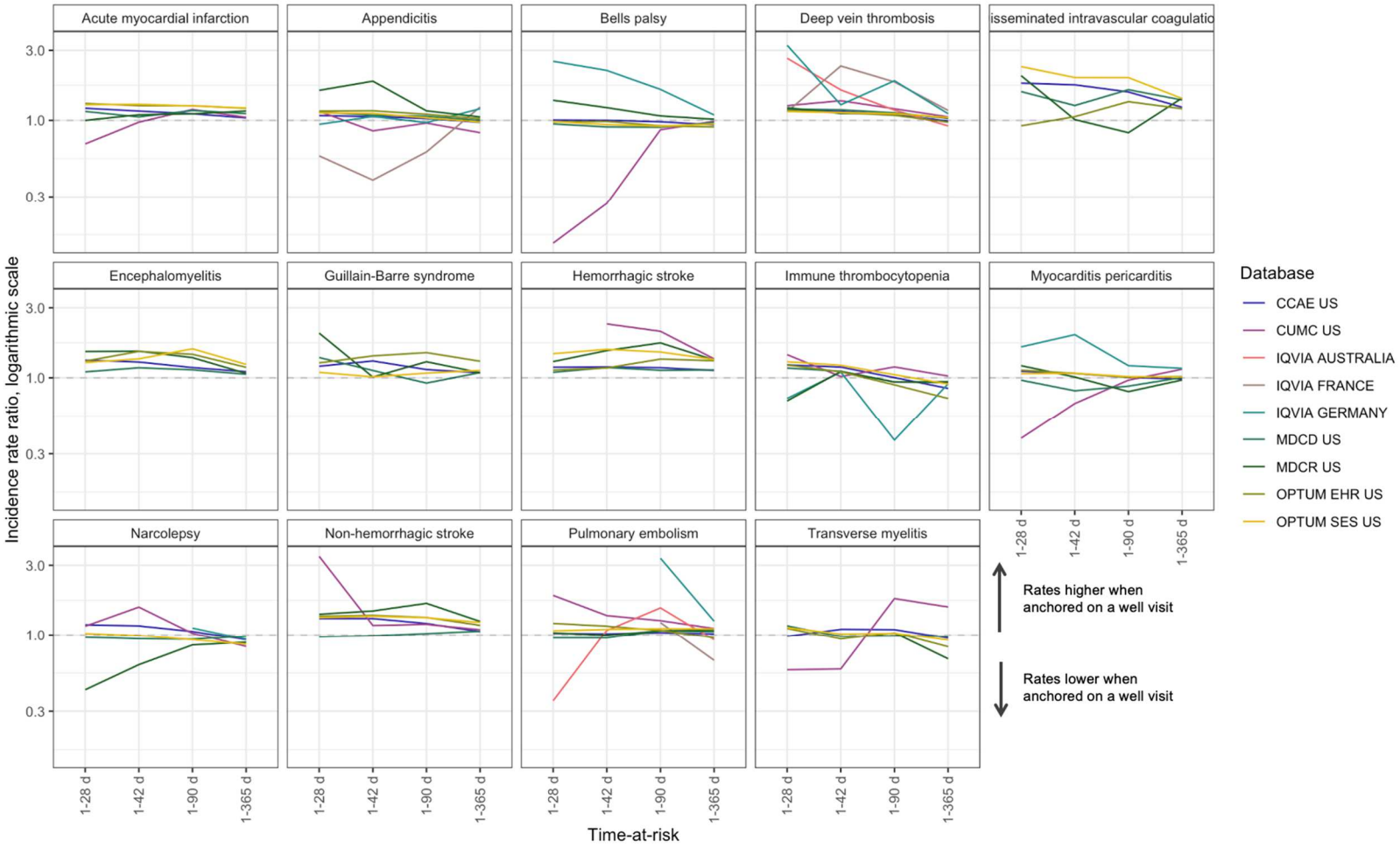


Appendix 4.10 Pooled age-adjusted incidence rate ratios (incidence rate of outcomes when entering the cohort on a well visit versus on January 1st in patients with a well visit in the next year) from meta-analyses, IRR and 95% CI.

Outcome	Time at risk 0-1 day	Time at risk 1-28 days	Time at risk 1-42 days	Time at risk 1-90 days	Time at risk 1-365 days
Acute myocardial infarction	1.14 (0.83-1.58)	1.24 (1.19-1.29)	1.19 (1.12-1.26)	1.18 (1.12-1.24)	1.13 (1.07-1.19)
Anaphylaxis	18.08 (9.47-34.5)	1.37 (1.3-1.44)	1.33 (1.27-1.38)	1.2 (1.15-1.25)	0.95 (0.92-0.97)
Appendicitis	3.44 (1.54-7.7)	1.11 (1.08-1.16)	1.11 (1.06-1.16)	1.06 (1.03-1.08)	1.01 (0.98-1.04)
Bell's palsy	16.71 (9.86-28.29)	0.98 (0.94-1.03)	0.96 (0.91-1.01)	0.93 (0.9-0.97)	0.93 (0.91-0.96)
Deep vein thrombosis	15.5 (9.29-25.86)	1.16 (1.13-1.19)	1.13 (1.11-1.15)	1.09 (1.07-1.12)	1.01 (0.98-1.04)
Disseminated intravascular coagulation	6.92 (3.19-14.97)	1.61 (1.2-2.17)	1.49 (1.18-1.86)	1.58 (1.34-1.86)	1.3 (1.23-1.38)
Encephalomyelitis	1.55 (0.61-3.91)	1.27 (1.05-1.52)	1.31 (1.13-1.52)	1.31 (1.16-1.48)	1.14 (1.08-1.2)
Guillain-Barre syndrome	2.68 (0.91-7.88)	1.2 (0.93-1.54)	1.18 (0.96-1.46)	1.14 (0.99-1.32)	1.13 (1.04-1.22)
Hemorrhagic stroke	0.88 (0.57-1.37)	1.23 (1.08-1.41)	1.29 (1.11-1.5)	1.32 (1.16-1.5)	1.22 (1.15-1.3)
Immune thrombocytopenia	30.81 (14.03-67.65)	1.22 (1.14-1.31)	1.16 (1.09-1.23)	0.97 (0.91-1.03)	0.86 (0.78-0.95)
Myocarditis and pericarditis	10.03 (5.44-18.49)	1.08 (1.01-1.17)	1.05 (0.98-1.11)	0.99 (0.95-1.04)	1 (0.98-1.02)
Narcolepsy	29.85 (16.28-54.72)	1.09 (1-1.19)	1.06 (0.97-1.15)	1 (0.95-1.05)	0.92 (0.88-0.96)
Non-hemorrhagic stroke	1.2 (0.59-2.42)	1.25 (1.15-1.37)	1.27 (1.16-1.38)	1.25 (1.15-1.36)	1.13 (1.08-1.19)

Pulmonary embolism	12.32 (7.35-20.65)	1.07 (1.01-1.13)	1.06 (1.02-1.1)	1.06 (1.04-1.09)	1.03 (0.98-1.1)
Transverse myelitis	15.59 (6.81-35.66)	1.05 (0.87-1.25)	1.02 (0.88-1.18)	1.05 (0.95-1.17)	0.92 (0.87-0.97)
All	6.32 (3.1-12.89)	1.17 (1.11-1.23)	1.15 (1.09-1.2)	1.11 (1.06-1.16)	1.04 (0.99-1.09)

Appendix 4.11 Incidence rate ratio of incidence rate of outcomes when entering the cohort on a well visit versus entering on January 1st in patients with a well visit in the next year, time-at-risk 1-28 days, 1-42 days, 1-90 days and 1-365 days time-at-risk.

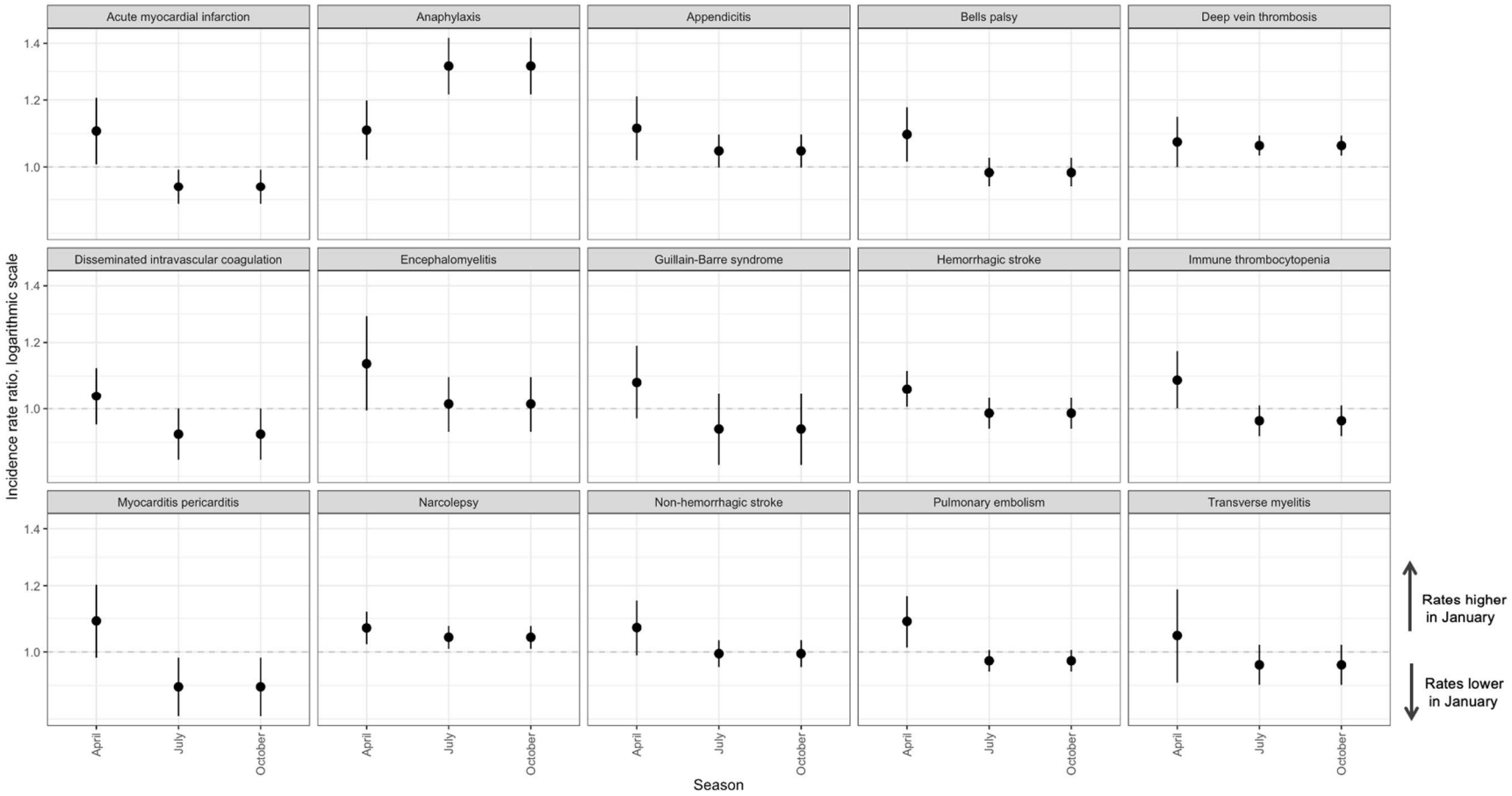


Appendix 4.12 Seasonal trends. Incidence rate ratios of incidence rates for quarter 2 (April - June), quarter 3 (July - September) and quarter 4 to quarter 1 (January – March) of 2017-2019 from meta-analysis, IRR and 95% CI.

Outcome	Quarter 2 versus Quarter 1	Quarter 3 versus Quarter 1	Quarter 4 versus Quarter 1
Acute myocardial infarction	1.1 (1.01-1.21)	0.95 (0.9-0.99)	0.95 (0.9-0.99)
Anaphylaxis	1.11 (1.02-1.2)	1.32 (1.22-1.42)	1.32 (1.22-1.42)
Appendicitis	1.11 (1.02-1.21)	1.04 (1-1.09)	1.04 (1-1.09)
Bell's palsy	1.09 (1.01-1.18)	0.99 (0.95-1.03)	0.99 (0.95-1.03)
Deep vein thrombosis	1.07 (1-1.15)	1.06 (1.03-1.09)	1.06 (1.03-1.09)
Disseminated intravascular coagulation	1.04 (0.96-1.12)	0.93 (0.87-1)	0.93 (0.87-1)
Encephalomyelitis	1.13 (1-1.29)	1.01 (0.94-1.09)	1.01 (0.94-1.09)
Guillain-Barre syndrome	1.08 (0.97-1.19)	0.95 (0.86-1.04)	0.95 (0.86-1.04)
Hemorrhagic stroke	1.06 (1.01-1.11)	0.99 (0.95-1.03)	0.99 (0.95-1.03)
Immune thrombocytopenia	1.08 (1-1.17)	0.97 (0.93-1.01)	0.97 (0.93-1.01)
Myocarditis and pericarditis	1.09 (0.98-1.2)	0.91 (0.84-0.98)	0.91 (0.84-0.98)
Narcolepsy	1.07 (1.02-1.12)	1.04 (1.01-1.07)	1.04 (1.01-1.07)
Non-hemorrhagic stroke	1.07 (0.99-1.15)	1 (0.96-1.03)	1 (0.96-1.03)

Pulmonary embolism	1.09 (1.01-1.17)	0.98 (0.95-1.01)	0.98 (0.95-1.01)
Transverse myelitis	1.05 (0.92-1.19)	0.97 (0.91-1.02)	0.97 (0.91-1.02)
All	1.08 (1.06-1.1)	1 (0.97-1.03)	1 (0.97-1.03)

Appendix 4.13 Comparison of incidence rates for quarter 2 (April - June), quarter 3 (July - September) and quarter 4 to quarter 1 (January – March) of 2017-2019 from meta-analysis, IRR and 95% CI.

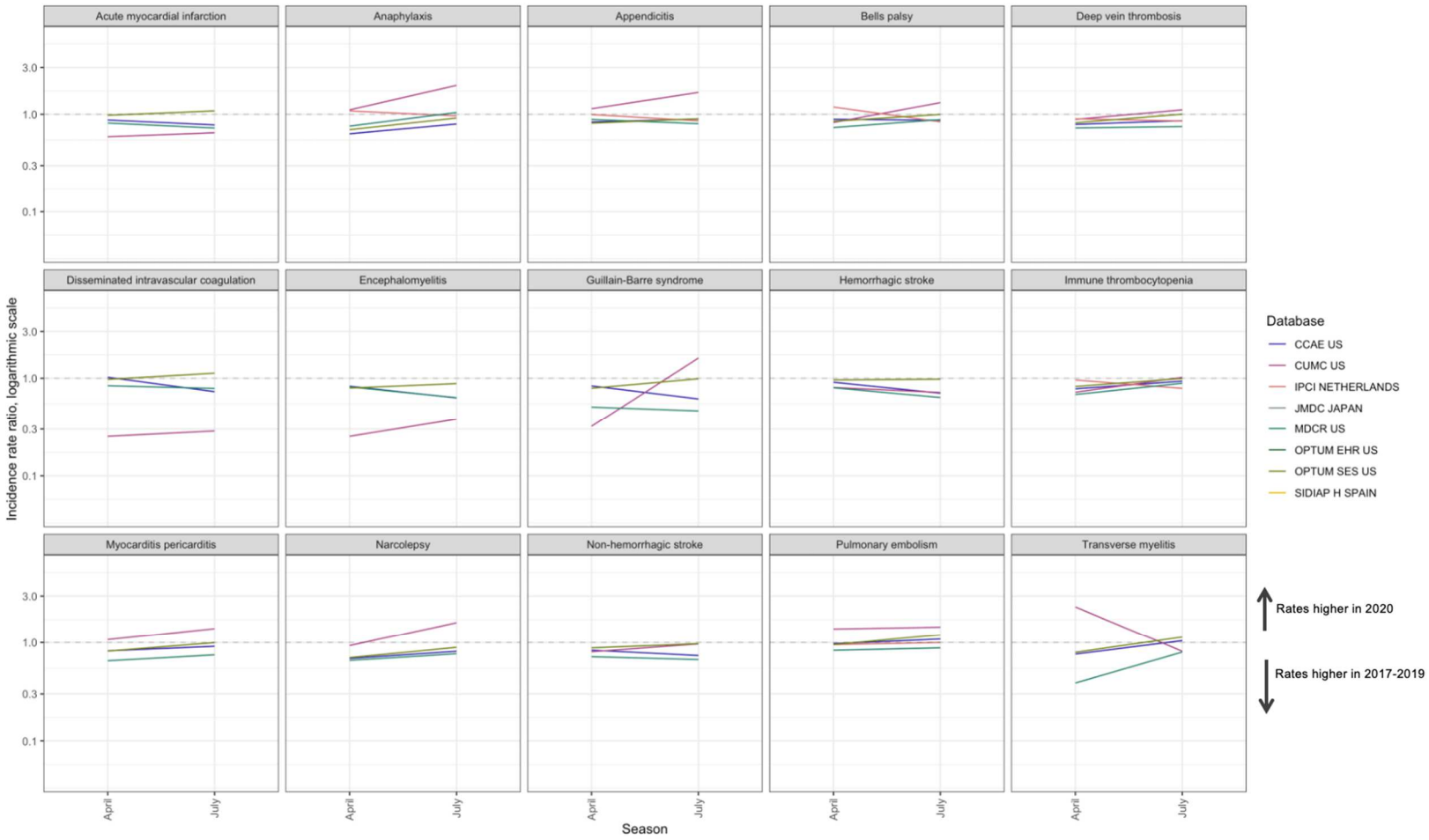


Appendix 4.14 Incidence rate ratios of incidence rates for Quarter 2 and 3 of 2020 versus Quarter 2 and 3 of 2017 – 2019 from meta-analysis, IRR and 95% CI.

Outcome	Quarter 2 2020 versus Quarter 2 2017-2019	Quarter 3 2020 versus Quarter 3 2017-2019
Acute myocardial infarction	0.95 (0.9-0.99)	0.92 (0.77-1.09)
Anaphylaxis	1.32 (1.22-1.42)	1.28 (0.93-1.76)
Appendicitis	1.04 (1-1.09)	1.16 (0.91-1.49)
Bell's palsy	0.99 (0.95-1.03)	1.12 (0.9-1.41)
Deep vein thrombosis	1.06 (1.03-1.09)	1.04 (0.86-1.24)
Disseminated intravascular coagulation	0.93 (0.87-1)	0.81 (0.63-1.04)
Encephalomyelitis	1.01 (0.94-1.09)	0.84 (0.63-1.11)
Guillain-Barre syndrome	0.95 (0.86-1.04)	0.83 (0.55-1.26)
Hemorrhagic stroke	0.99 (0.95-1.03)	0.85 (0.7-1.04)
Immune thrombocytopenia	0.97 (0.93-1.01)	1.08 (0.89-1.3)
Myocarditis and pericarditis	0.91 (0.84-0.98)	1.18 (0.86-1.62)
Narcolepsy	1.04 (1.01-1.07)	1.15 (0.86-1.52)
Non-hemorrhagic stroke	1 (0.96-1.03)	0.95 (0.77-1.17)

Pulmonary embolism	0.98 (0.95-1.01)	1.25 (1.03-1.52)
Transverse myelitis	0.97 (0.91-1.02)	1.09 (0.93-1.27)
All	1 (0.97-1.03)	1.03 (0.96-1.1)

Appendix 4.15 Comparison of incidence rates in Q2, Q3 of 2017 – 2019 (pre-COVID-19 pandemic) versus corresponding quarters in 2020 (COVID-19 pandemic), incidence rate ratios.

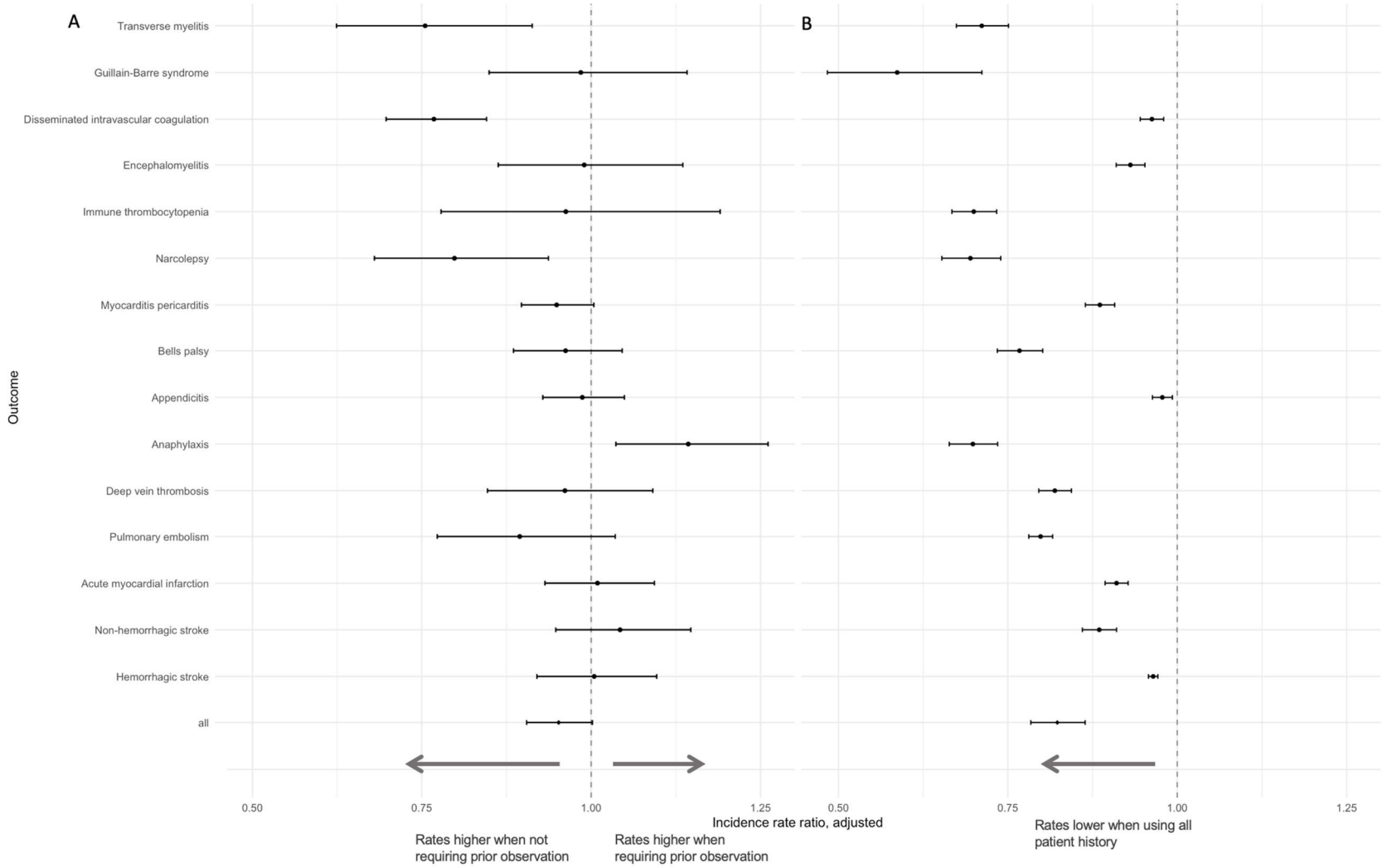


Appendix 4.16 Pooled age-adjusted incidence rate ratios for comparing clean windows and prior observation, from meta-analyses, IRR and 95% CI.

Outcome	First ever conditions in patient history versus first occurrence in a given window	Requirement to have a year of prior observation versus no prior observation requirement
Acute myocardial infarction	0.91 (0.89-0.93)	1.01 (0.93-1.09)
Anaphylaxis	0.7 (0.66-0.73)	1.14 (1.04-1.26)
Appendicitis	0.98 (0.96-0.99)	0.99 (0.93-1.05)
Bell's palsy	0.77 (0.73-0.8)	0.96 (0.89-1.05)
Deep vein thrombosis	0.82 (0.8-0.84)	0.96 (0.85-1.09)
Disseminated intravascular coagulation	0.96 (0.95-0.98)	0.77 (0.7-0.85)
Encephalomyelitis	0.93 (0.91-0.95)	0.99 (0.86-1.14)
Guillain-Barre syndrome	0.59 (0.48-0.71)	0.98 (0.85-1.14)
Hemorrhagic stroke	0.96 (0.96-0.97)	1 (0.92-1.1)
Immune thrombocytopenia	0.7 (0.67-0.73)	0.96 (0.78-1.19)
Myocarditis and pericarditis	0.89 (0.86-0.91)	0.95 (0.9-1)
Narcolepsy	0.69 (0.65-0.74)	0.8 (0.68-0.94)

Non-hemorrhagic stroke	0.88 (0.86-0.91)	1.04 (0.95-1.15)
Pulmonary embolism	0.8 (0.78-0.82)	0.89 (0.77-1.04)
Transverse myelitis	0.71 (0.67-0.75)	0.75 (0.62-0.91)
All	0.82 (0.78-0.86)	0.95 (0.9-1)

Appendix 4.17 Comparison of prior observation (A) and clean windows (B) , from meta-analyses, IRR and 95% CI.



Appendix 4.18 Negative controls for COVID-19 vaccine effectiveness study.

SNOMED concept id	SNOMED concept name
438945	Accidental poisoning by benzodiazepine-based tranquilizer
434455	Acquired claw toes
316211	Acquired spondylolisthesis
201612	Alcoholic liver damage
438730	Alkalosis
441258	Anemia in neoplastic disease
432513	Animal bite wound
4171556	Ankle ulcer
4098292	Antiphospholipid syndrome
77650	Aseptic necrosis of bone
4239873	Benign neoplasm of ciliary body
23731	Benign neoplasm of larynx
199764	Benign neoplasm of ovary
195500	Benign neoplasm of uterus

4145627	Biliary calculus
4108471	Burn of digit of hand
75121	Burn of lower leg
4284982	Calculus of bile duct without obstruction
434327	Cannabis abuse
78497	Cellulitis and abscess of toe
4001454	Cervical spine ankylosis
4068241	Chronic instability of knee
195596	Chronic pancreatitis
4206338	Chronic salpingitis
4058397	Claustrophobia
74816	Contusion of toe
73302	Curvature of spine
4151134	Cyst of pancreas
77638	Displacement of intervertebral disc without myelopathy

195864	Diverticulum of bladder
201346	Edema of penis
200461	Endometriosis of uterus
377877	Esotropia
193530	Follicular cyst of ovary
4094822	Foreign body in respiratory tract
443421	Gallbladder and bile duct calculi
4299408	Gouty tophus
135215	Hashimoto thyroiditis
442190	Hemorrhage of colon
43020475	High risk heterosexual behavior
194149	Hirschsprung's disease
443204	Human ehrlichiosis
4226238	Hyperosmolar coma due to diabetes mellitus
4032787	Hyperosmolarity

197032	Hyperplasia of prostate
140362	Hypoparathyroidism
435371	Hypothermia
138690	Infestation by Pediculus
4152376	Intentional self poisoning
192953	Intestinal adhesions with obstruction
196347	Intestinal parasitism
137977	Jaundice
317510	Leukemia
765053	Lump in right breast
378165	Nystagmus
434085	Obstruction of duodenum
4147016	Open wound of buttock
4129404	Open wound of upper arm
438120	Opioid dependence

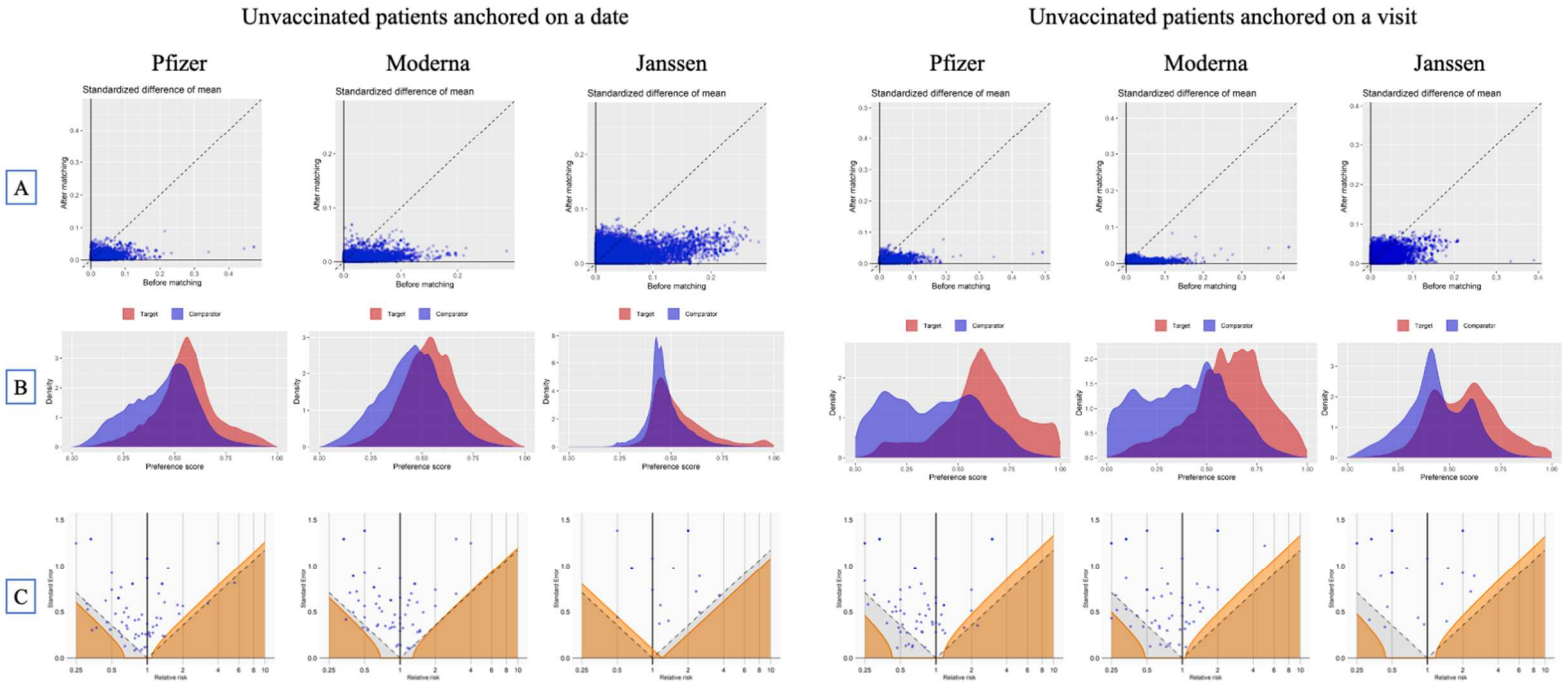
75924	Osteodystrophy
432594	Osteomalacia
30365	Panhypopituitarism
4108371	Peripheral gangrene
440367	Plasmacytosis
439233	Poisoning by antidiabetic agent
442149	Poisoning by bee sting
4314086	Poisoning due to sting of ant
4147660	Postural kyphosis
434319	Premature ejaculation
199754	Primary malignant neoplasm of pancreas
4311499	Primary malignant neoplasm of respiratory tract
436635	Primary malignant neoplasm of sigmoid colon
196044	Primary malignant neoplasm of stomach
433716	Primary malignant neoplasm of testis

133424	Primary malignant neoplasm of thyroid gland
194997	Prostatitis
80286	Prosthetic joint loosening
443274	Psychostimulant dependence
314962	Raynaud's disease
37018294	Residual osteitis
4288241	Salmonella enterica subspecies arizonae infection
45757269	Sclerosing mesenteritis
74722	Secondary localized osteoarthrosis of pelvic region
200348	Secondary malignant neoplasm of large intestine
43020446	Sedative withdrawal
74194	Sprain of spinal ligament
4194207	Tailor's bunion
193521	Tropical sprue
40482801	Type II diabetes mellitus uncontrolled

74719	Ulcer of foot
196625	Viral hepatitis A without hepatic coma
197494	Viral hepatitis C
4284533	Vitamin D-dependent rickets

Appendix 4.19 Diagnostics for the effectiveness study comparing the cohort vaccinated with at least one dose of Pfizer, Moderna or Janssen COVID-19 vaccines and unvaccinated cohort anchored on a date or on a visit: (A) covariate balance before and after propensity score matching, (B) preference score balance and (C) effect of negative control calibration displaying effect estimate and standard error.

In (A), each dot represents the standardized difference of the means for a single covariate before and after stratification on the propensity score. In (C), each blue dot is a negative control. The area below the dashed line indicates estimates with $p<0.05$ and the orange area indicates estimates with calibrated $p<0.05$.

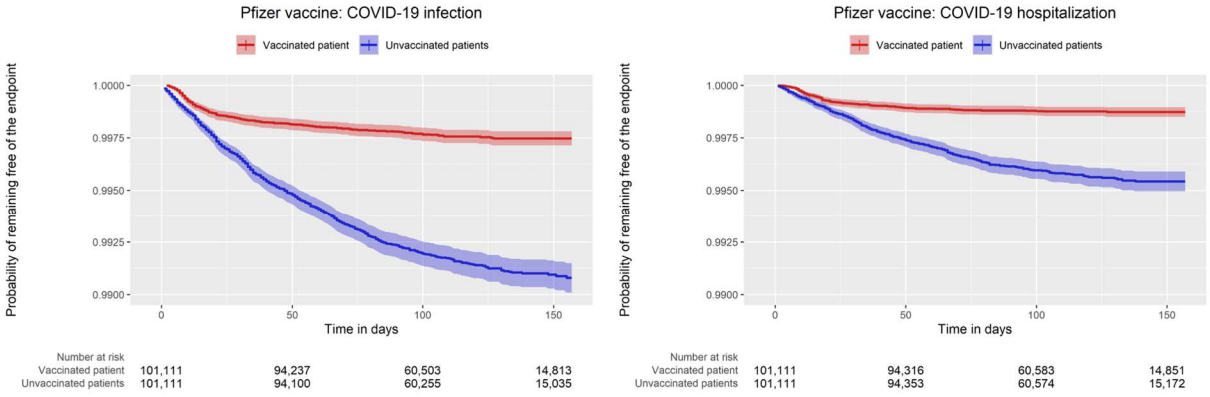

Unvaccinated patients anchored on a date
Unvaccinated patients anchored on a visit
Pfizer
Moderna
Janssen
Pfizer
Moderna
Janssen
A
B
C
Standardized difference of mean
After matching
Before matching
Target
Comparator
Density
Preference score
Standard Error
Relative risk

Appendix 4.20 Summary of manual chart review of COVID-19 infection cases during week 1 after the index date, patients vaccinated with mRNA vaccines and unvaccinated patients.

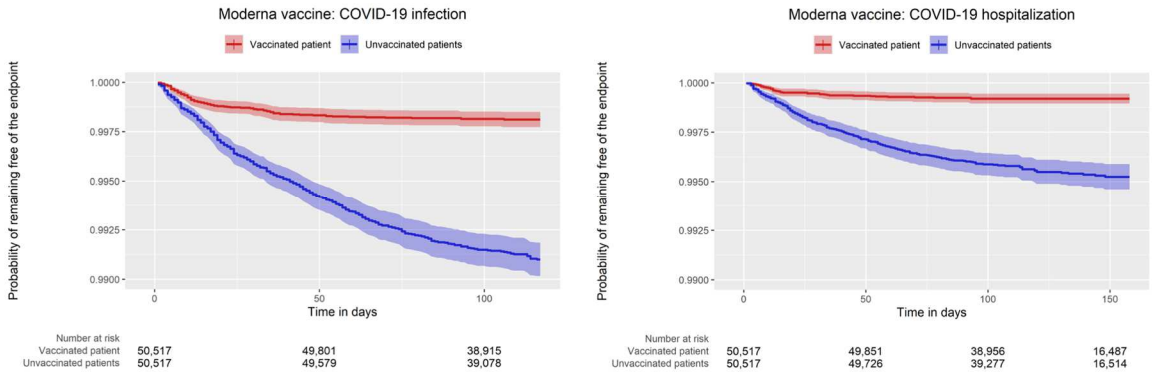
	Pfizer- BioNTech	**Moderna**	**Pfizer- BioNTech and Moderna**	**Unvaccinated patients**
Total	36	25	61	28
Average age	65.0	67.8	65.8	58.0
COVID-19 symptoms				
Severe	14 (39%)	7 (28 %)	21 (34%)	6 (21%)
Mild	18 (50%)	11 (44%)	29 (48%)	11 (39%)
Asymptomatic	4 (11%)	7 (28%)	11 (18%)	11 (39%)
Reason for coming for initial healthcare encounter				
COVID-19 symptoms	18 (50%)	12 (48%)	30 (49%)	18 (64%)
Exposure to COVID-19	3 (8%)	4 (16%)	7 (12%)	6 (21%)
For other reason (co-morbidities, procedures etc.)	15 (42%)	9 (36%)	24 (39%)	4 (14%)
Type of initial healthcare encounter				

Telehealth/phone	5 (14%)	6 (24%)	11 (18%)	3 (11%)
Test only	3 (8%)	2 (8%)	5 (8%)	6 (21%)
Outpatient visit	4 (11%)	3 (12%)	7 (12%)	1 (4%)
Emergency room or inpatient visit	24 (67%)	14 (56%)	38 (62%)	18 (64%)

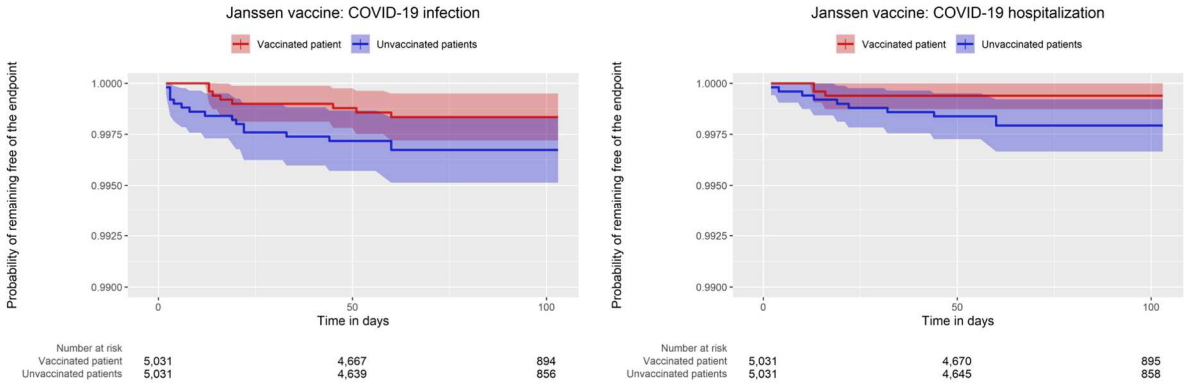
Appendix 4.21 Kaplan-Meier curves for effectiveness of COVID-19 Pfizer-BioNTech vaccine for time-at-risk of 1 day – 365 days after the first dose compared to the unvaccinated patients residing in New York City.



Appendix 4.22 Kaplan-Meier curves for effectiveness of COVID-19 Moderna vaccine for time-at-risk of 1 day – 365 days after the first dose compared to the unvaccinated patients residing in New York City.



Appendix 4.23 Kaplan-Meier curves for effectiveness of COVID-19 Janssen vaccine for time-at-risk of 1 day – 365 days after the first dose compared to the unvaccinated patients residing in New York City.



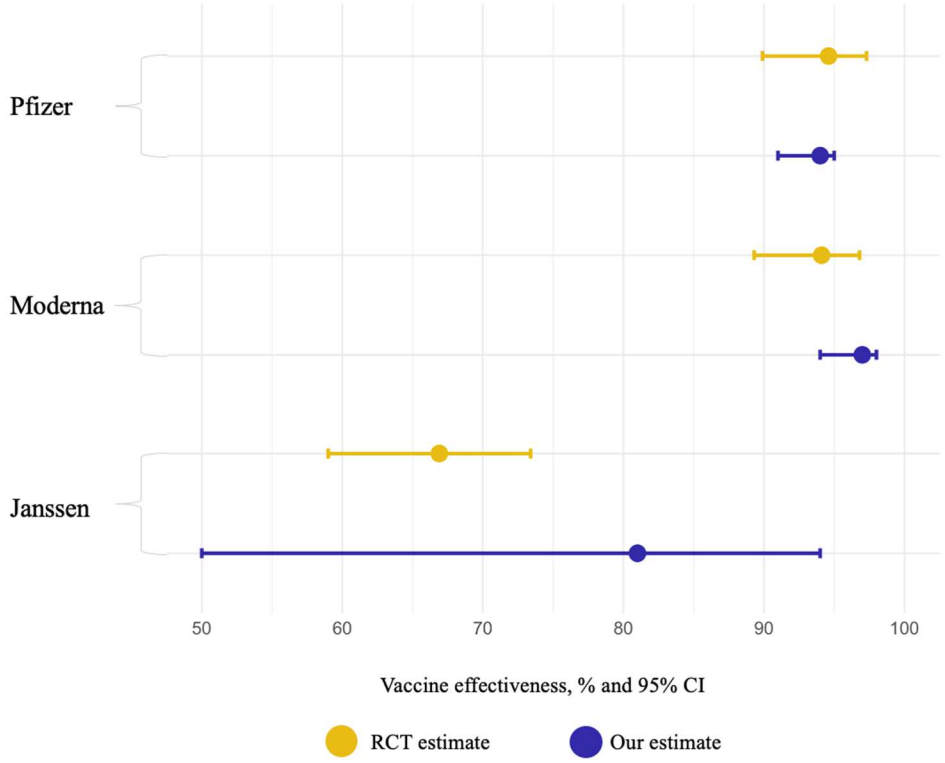
Appendix 4.24 Estimates for long-term effectiveness of COVID-19 vaccines for time-at-risk of 1 day – 365 days after the first dose in the vaccinated patients without prior COVID-19 infection compared to unvaccinated patients residing in NYC. VE – vaccine effectiveness.

	COVID-19 infection		**COVID-19 hospitalization**		**COVID-19 positive test only**		**COVID-19 positive test only hospitalization**	
	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value
Pfizer-BioNTech	42 (37 – 47)	<0.01	63 (56 – 70)	<0.01	71 (66 – 75)	<0.01	69 (62 – 75)	<0.01
Moderna	54 (48 – 60)	<0.01	76 (69 – 82)	<0.01	78 (73 – 83)	<0.01	81 (74 – 87)	<0.01
Janssen	24 (0 – 55)	0.31	64 (0.1 – 1.06)	0.09	53 (0 – 82)	0.1	70 (2 – 93)	0.08

Appendix 4.25 Estimates for effectiveness of COVID-19 vaccines for time-at-risk of 1 day – 365 days after full vaccination in fully vaccinated patients without prior COVID-19 infection compared to unvaccinated patients residing in NYC. VE – vaccine effectiveness.

	COVID-19 positive test only		**COVID-19 positive test only hospitalization**		**COVID-19 infection**		**COVID-19 hospitalization**	
	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value
Pfizer-BioNTech	94 (91 – 95)	<0.01	95 (92 – 97)	<0.01	70 (66 – 74)	<0.01	88 (84 – 92)	<0.01
Moderna	97 (94 – 98)	<0.01	96 (92 – 99)	<0.01	72 (66 – 77)	<0.01	92 (87 – 95)	<0.01
Janssen	81 (50 – 94)	<0.01	92 (58 – 100)	0.03	55 (23 – 75)	0.01	87 (56 – 98)	0.01

Appendix 4.26 Comparison of the effectiveness estimates in fully vaccinated patients obtained in our study and those from the randomized clinical trials of the corresponding vaccines.



Appendix 4.27 Estimates for effectiveness of COVID-19 vaccines for time-at-risk of 1 day – 365 days after the first dose in the vaccinated patients with or without prior COVID-19 infection compared to unvaccinated patients residing in NYC. VE – vaccine effectiveness.

	COVID-19 infection		**COVID-19 hospitalization**		**COVID-19 positive test only**		**COVID-19 positive test only hospitalization**	
	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value	VE, % (95% CI)	P-value
Pfizer- BioNTech	43 (38 – 48)	<0.01	64 (57 – 70)	<0.01	71 (66 – 75)	<0.01	71 (64 – 76)	<0.01
Moderna	51 (45 – 57)	<0.01	71 (63 – 78)	<0.01	76 (71 – 81)	<0.01	81 (73 – 86)	<0.01
Janssen	15 (0 – 49)	0.52	60 (2 – 86)	0.06	45 (0 – 75)	0.12	63 (0 – 90)	0.09

Appendix for Chapter 5

Appendix 5.1 Distribution of primary source of evidence, number of questions and types of questions depending on primary working settings, specialization and academic rank.

		Primary setting		**Specialization**		**Academic rank**		
		Inpatient	Outpatient	Primary	Specialty	1*	2	3
Primary source of evidence	Commercial tool	8 (38%)	6 (86%)	7 (70%)	7 (39%)	5 (71%)	5 (41.7%)	4 (44.4%)
	Guidelines	2 (9.6%)	1 (14%)	2 (20%)	1 (5.5%)	1 (14.5%)	1 (8.3%)	1 (11.2%)
	PubMed	9 (42.9%)	0 (0%)	0 (0%)	9 (50%)	0 (0%)	5 (41.7%)	4 (44.4%)
	Other literature	1 (4.8%)	0 (0%)	0 (0%)	1 (5.5%)	1 (14.5%)	0 (0%)	0 (0%)
	None	1 (4.8%)	0 (0%)	1 (10%)	0 (0%)	0 (0%)	1 (8.3%)	0 (0%)
Average number of questions		4.9±2.2	2.7±1.1	3.4±1.6	4.8±2.3	2.9±1.1	4.2±1.8	5.7±2.6
Average number	Diagnosis	0.4±0.6	0.9±0.7	0.7±0.7	0.4±0.6	0.5±0.5	0.3±0.5	0.8±0.8
	Treatment	4.2±2.0	1.4±0.9	2.3±1.8	4.2±2.0	2.3±0.9	3.7±1.8	4.2±2.9

of questions per type	Quality of care	0.3±0.6	0.4±0.8	0.4±0.7	0.3±0.5	0.1±0.4	0.2±0.4	0.7±0.8

* 1 – none , 2 – Assistant Professor, 3 – Associate Professor

Appendix 5.2 Examples of questions within the identified themes, thematic analysis of clinicians' immediate information needs.

Category	Example	Details
Safety & Adverse Events	What are the risk factors for vancomycin-induced kidney injury?	
	How implantable defibrillator impacts the risk of sudden cardiac death in patients with tetralogy of Fallot?	
Comparative Effectiveness	For an elderly woman with a drug-eluting stent, atrial fibrillation and a pacemaker, what should	

	we prescribe: clopidogrel, aspirin, warfarin or some combinations?	
	How well does metoprolol control heart rate in patients with atrial fibrillation versus carvedilol?	
Necessity of Treatment	What are the rules of prescribing antibiotics in patients with peritoneal dialysis catheter?	"A patient on peritoneal dialysis with catheter for a year asked if she needs an antibiotic prophylaxis for a dental procedure. I wasn't sure what to answer"
	Should we prescribe vitamin D in children knowing that compliance may be low?	"We know that patients do not take vitamin D tablets as they need to take them once a week. They just forget. Knowing that, do we even need to prescribe it if they won't take it?"
Indication/ Contraindication	Can Lovenox be prescribed in patients with low kidney function?	

	Should a diabetic patient on ACE inhibitors, diuretics and SGLT2 inhibitors be taken off diuretics as SGLT2 inhibitors act as diuretics?	
Dose Ranges	When guidelines state dose range, which dose from the range should we pick?	
	What is the minimal dosage of desmopressin in sick children, so that this dose does not cause harmful adverse effects but still has therapeutic effects?	"Giving all the confounding factors [that he is tiny, that he is sick, that his absorption is not reliable] what is the tiniest dose that we can give and with that dose what is the maximum duration of action of dose in this child?"
Dose Adjustment	How do we adjust the dose of Lovenox in obese patients to avoid drug toxicity?	"We see patients coming to us with a prescription of 100 milligrams. It is a lot, you know, and we would never suggest that."

Duration of Therapy	In patients on steroids for acute or chronic adrenal insufficiency, asthma or other chronic conditions, what is the optimal dose, how long should a patient stay on steroids and when to cancel?	
	How long should we continue antibiotic therapy for complicated UTI for?	
Test Comparative Effectiveness	PPD versus QuantiFERON: what is better?	
Test Choice	How do we properly diagnose ovarian tumors in children?	"I think, there is a differential [diagnosis] for ovarian tumors of childhood: granulomas, cell cysts, bad granulomas. And the tumor markers are often non-informative: alpha-fetoprotein, beta-HCG is not

		informative. Sometimes hormones are informative, but not always.”
	What diagnostic workup is indicated for a patient with duodenal villus atrophy but negative celiac disease antibodies?	
Test Interpretation	How do we interpret Synachten test in patients on long-term steroid therapy?	“We have patients that have been on steroids for months or even years, and we don’t really know if they need them. Then they stop the medications themselves and live just fine. We perform ACTH [adrenocorticotropic hormone] test, but different doctors have different opinions on the results: one says it should drop 10 points, another - 10 points but not less than 18 and so on.”
Test Frequency	In child with hyperthyroidism treated with methimazole and dexamethasone, how frequently should we measure T3 and T4?	“We got a patient with hypothyroidism due to maternal Graves’ disease, we put him on

		methimazole and steroids. Baby's T3 and T4 initially improved, we cut back dexamethasone. Then they went up. We increased the dose of methimazole, baby was stable, but the labs still went up. How frequently should the labs be checked for such a patient? In this case, the daily checks might have fooled us."
Symptoms and Syndromes	Can obesity inflammation alone get erythrocyte sedimentation rate up to a very high level?	"There have been cases where a high BMI [body mass index] patient has a persistent white blood cell count of 12-13 or those who have ESRs [erythrocyte sedimentation rate] over 70, cases where we may feel obligated to do a major workup and still to be left with no answers. I asked my colleague last month, and he feels he doesn't seen values this high just from plain obesity."
	Figure out the patterns and patient trajectories for somatic disorder.	"We have these patients [with somatic disorder] all the time. They come and go, and it takes years for them to be

		diagnosed. They complain about back pain, headaches, heart problems and we can't let them go without a workup or just say that it's nothing. And physicians don't want to deal with them because they are scared by the disorder and don't know what to do."
Screening	How should patients be screened for dementia?	"It [dementia] often is confused with depression especially if a person has a history of depression. We need this information to properly educate primary physicians on how to take care of such patients."
Prognosis	What is the risk of epilepsy after a mild brain injury and how can this risk be reduced?	"A patient was skiing and had a concussion, so he called me and asked if he should stay in bed or if he could take bath. We advised him not to take it as it might provoke seizures. But we might have sent him to the hospital if we knew that the risk was substantial."

Social Work/Ancillary Services	What is the evidence on the impact of home care on patient outcomes in patients receiving mechanical ventilation?	

Appendix 5.3 Search strategy for scoping review of clinical decision support tools that generate new evidence

1. PubMed

("evidence-based medicine"[All Fields] OR "evidence-based practice"[All Fields] OR "evidence" [All Fields] OR EBM[All Fields] OR evidence[All Fields]) AND ("health services"[Title/Abstract] OR "decision support systems"[Title/Abstract] OR "prototype"[Title/Abstract] or "medical information systems"[Title/Abstract] OR "computer-assisted decision making"[Title/Abstract] OR "clinical decision support"[Title/Abstract] OR software[Title/Abstract] OR "decision support systems, clinical"[MeSH Terms] OR "Decision Making, Computer-Assisted"[Mesh:noexp]) AND ("comorbidity"[tiab] OR "personalized"[tiab] OR "precision"[tiab] OR "complicated"[tiab] OR "multiple conditions"[tiab] OR "complex"[tiab] OR "similar patient"[tiab] OR "patient like mine"[tiab] OR "patient-like-mine"[tiab] OR "patient-related question"[tiab] OR "patient-related questions"[tiab] OR bedside[tiab] OR "point-of-care"[tiab] OR "point of care"[tiab] OR "real-time"[tiab] OR "real time"[tiab] OR just-in-time[tiab]).

2. IEEE Xplore

("electronic health record" OR EMR OR "electronic medical record" OR EHR OR “health record systems”) AND ("evidence-based medicine" OR "evidence-based practice" OR "evidence" OR EBM OR evidence) AND ("health services" OR "decision support systems” OR "prototype" or "medical information systems" OR "computer-assisted decision making" OR "clinical decision support" OR software) AND ("comorbidity" OR "personalized" OR "precision" OR "complicated" OR "multiple conditions" OR "complex" OR "similar patient" OR “patient like mine” OR “patient-like-mine” OR "patient-related question" OR "patient-related questions" OR bedside OR “point-of-care” OR “point of care” OR “real-time” OR “real time” OR just-in-time).

3. Embase

(‘electronic health records’:ti,ab,kw OR EHR:ti,ab,kw) AND (‘evidence-based medicine’:ti,ab,kw OR ‘evidence-based practice’:ti,ab,kw OR ‘evidence’:ti,ab,kw OR EBM:ti,ab,kw) AND (‘health services’:ti,ab,kw OR ‘decision support systems’:ti,ab,kw OR ‘prototype’:ti,ab,kw or ‘medical information systems’:ti,ab,kw OR ‘clinical decision support’:ti,ab,kw OR ‘medical information systems’:ti,ab,kw OR ‘computer-assisted decision making’:ti,ab,kw) AND (‘comorbidity’:ti,ab,kw OR ‘personalized’:ti,ab,kw OR ‘precision’:ti,ab,kw OR ‘complicated’:ti,ab,kw OR ‘multiple conditions’:ti,ab,kw OR ‘complex’:ti,ab,kw OR ‘similar patient’:ti,ab,kw OR “patient like mine”:ti,ab,kw OR “patient-like-mine”:ti,ab,kw OR "patient-related question":ti,ab,kw OR "patient-related

questions":ti,ab,kw OR bedside:ti,ab,kw OR “point-of-care”:ti,ab,kw OR “point of care”:ti,ab,kw OR “real-time”:ti,ab,kw OR “real time”:ti,ab,kw OR just-in-time:ti,ab,kw)

4. Proquest

Anywhere except full text: ("electronic health record" OR EMR OR "electronic medical record" OR EHR OR “health record systems”) AND ("evidence-based medicine" OR "evidence-based practice" OR "evidence" OR EBM OR evidence) AND ("health services" OR "decision support systems” OR "prototype" or "medical information systems" OR "computer-assisted decision making" OR "clinical decision support" OR software) AND ("comorbidity" OR "personalized" OR "precision" OR "complicated" OR "multiple conditions" OR "complex" OR "similar patient" OR “patient like mine” OR “patient-like-mine” OR "patient-related question" OR "patient-related questions" OR bedside OR “point-of-care” OR “point of care” OR “real-time” OR “real time” OR just-in-time). Document type restricted to : Article, Dissertation/Thesis, Conference Proceeding, Evidence Based Healthcare

Appendix 5.4 The main characteristics of the tools selected for this study.

Name	**Site of implementation**	**Primary Focus**	**Area**	**Stage**	**Evidence Mediator**	**Evaluation**	**Evaluation details**
Visual non-analytics-based data-driven tools							

Plaisant et al. ("PatternFinder"), 2006	One site (Washington Hospital Center, USA)	Clinical care, research	Unrestricted	Prototype	None	Functional testing	Case studies: contrast nephropathy, thrombocytopenia Methods: interview Participants: three emergency room clinicians
Perer et al. ("CareFlow"), 2013	One site (Watson Research Center, USA)	Clinical care	Potentially unrestricted	Prototype	None	Not reported	Case study: congestive heart failure
Li et al. ("Patient-like-mine"), 2015	One site (Mayo clinic, USA)	Clinical care	Surgery	Prototype	None	Not reported	
Zhang et al. ("CaVa"), 2015	One site (Watson Research Center, USA)	Research	Potentially unrestricted	Prototype	Research team	Functional testing	Case study: cardiology. Methods: interview Participants: one emergency room clinician

Happe et al. ("ePEPS") , 2018	Multiple sites (France, nationwide)	Clinical care, quality	Unrestricted	Prototype	None	Not reported	
Rogers et al. ("Composer"), 2019	One site (University of Utah, USA)	Clinical care, research	Surgery (orthopedic)	Prototype	None	Functional testing	Details not reported
Analytics-based data-driven tools							
Bernard et al., 2014	One site (University Hamburg-Eppendorf, Germany)	Research	Oncology (prostate cancer)	Implemented, actual use is not reported	None	Functional testing	Methods: interview Participants: six physicians
Bernard et al., 2019	One site (University of Leeds, UK)	Clinical care, research	Oncology (prostate cancer)	Prototype	None	Functional testing	Methods: Interview

							Participants: 14 (non-expert, visualization experts and medical experts)
Nan Cao et al., 2011	One site (University Hospital of North Norway, Norway)	Research	Unrestricted	Prototype	None	Functional and user acceptance testing	Case studies: disease distribution, similar patients Methods: Interview, performance metrics (task response time, task success rate) Participants: 30 students
Malik et al. ("CoCo"), 2015	One site (University of Maryland, USA)	Research	Unrestricted	Prototype	None	Functional testing	Case study: adherence to Advanced Trauma Life Support Protocol Methods: Multi-dimensional, long-term in-depth case study Participants: clinicians

Gallego et al. (“Green Button”), 2018	One site (Stanford Medical Center, USA)	Clinical care	Unrestricted	Implemented, pilot study in Stanford Medical Center	Research team	Feasibility study	Ongoing
Yu et al. ("Care Pathway Workbench"), 2014	One site (Watson Research Center, USA)	Clinical care	Potentially unrestricted	Prototype	None	Not reported	Case study: congestive heart failure
Mane et al. ("VisualDecisionLinc"), 2011	One site (Duke Medical Center, USA)	Clinical care, research	Psychiatry	Prototype	None	Not reported	Case study: major depressive disorder
Xia et al., 2017	One site (the First Affiliated Hospital, Xiamen	Clinical care	Internal medicine	Prototype, acute	None, but depends on the	Not reported	

	University, China)			coronary syndrome	study team		
Xia et al., 2019	One site (IBM research, China)	Clinical care	Internal medicine	Prototype, type II diabetes mellitus	None, but depends on the study team	Not reported	
Morrison et al., 2015	One site (University of Maryland, USA)	Clinical care	Oncology (lung cancer)	Prototype	None	Not reported	
Finlayson et al.("Melanoma Rapid Learning Utility"), 2016	Two cites (Vanderbilt-Ingram Cancer Center and	Clinical care, research	Oncology (melanoma)	Implemented, actual use is not reported	None	User acceptance testing	Methods: Interview, survey Participants: 13 clinicians

	Stanford Hospital, USA)						
Expert-based (knowledge-aggregative) tools							
Ellis et al. ("ViaOncology"), 2013	Multiple sites (40 University of Pittsburgh Medical Center affiliated sites)	Clinical care, quality	Oncology	Implemented, 150 specialists at the UPMC Cancer Center at the time of publication	None, but data supplied quarterly	Quality and patient outcomes	Methods: retrospective pretest–post-test study Participants: 172 patients with metastatic colorectal cancer Pathway adoption and cost of care
Hoverman et al. ("Level I Pathways"), 2008	Multiple sites (USA Oncology network practices)	Clinical care, quality	Oncology (evaluated colorectal,	Implemented, in use up to now in	None, but depends on the	Quality and patient outcomes	Methods: multiple prospective cohort studies Pathway adoption, cost of care, hospitalization rates

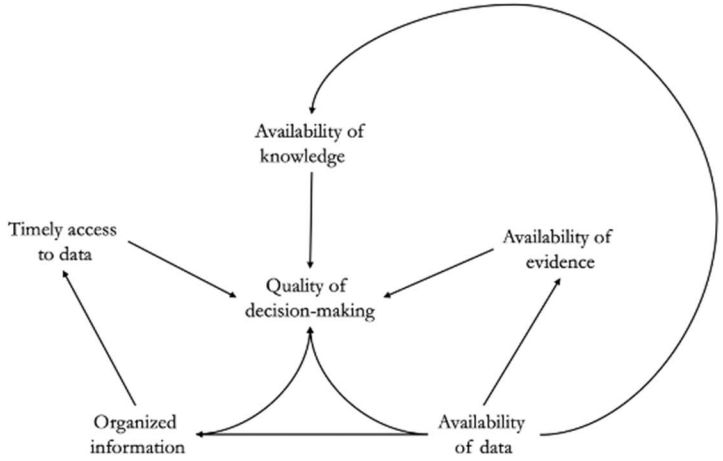
			lung, breast cancer)	affiliated centers	study team		a) prospective, nonrandomized cohort study Target group (on-pathway) – 168 and comparison - 53 patients with breast, lung cancer Outcomes: hospitalization rates and length of stay b) retrospective cohort study Target group (on-pathway) – 756 and comparison – 154 patients with colon cancer Outcomes: hospitalization rates, length of therapy, cost of care c) retrospective cohort study Target group (on-pathway) – 1,095 and comparison – 304 patients with colon cancer

							Outcomes: survival rate, cost of care
Feinberg et al. (“P4 Pathways”), 2012	Multiple sites (136 USA practices)	Clinical care, quality	Oncology (evaluated colorectal, lung, breast cancer)	Implemented, in use up to now in affiliated centers	None, but depends on the study team	Quality and patient outcomes	Methods: retrospective pretest–post-test study Participants: 4,713 patients with breast, lung, or colorectal cancer Pathway adoption and cost of care
Simon et al. ("Oncology Expert Advisor”), 2019	One site (MD Anderson Cancer Center, USA)	Clinical care, quality	Oncology (lung cancer)	Implemented, actual use is not reported	None, but depends on the study team	User acceptance testing	Details not reported
Karamlou et al., (“ROAD2H”), 2019	2 sites (China National Health Development	Clinical care, quality	COPD, CKD,	Implemented, actual use is not reported	None, but depends on the	User acceptance testing	Details not reported

	Research Center, University of Belgrade, Serbia)		potentially unrestricted		study team		
Henry et al. ("e-bipolar"), 2010	Multiple sites (France, nationwide)	Clinical care, quality	Psychiatry (bipolar disorder)	Implemented, actual use is not reported	None, but depends on the study team	Not reported	
Cook et al. ("MayoExpert"), 2017	Multiple sites (Mayo Clinic sites, USA)	Clinical care	Unrestricted	Implemented, at least for 2014-2015	None, but depends on the study team	Usage	Methods: retrospective longitudinal study Participants: data logs from 6700 clinicians with the access to the tool Tool adoption and use

* Evidence mediator – a person, team or organization that plays the primary role in evidence generation or synthesis.

Figure 5.5 Factors influencing the quality of decision-making through a causal-loop diagram.



Appendix 5.6 Clinical questions supplied to the Data Consult Service.

Characteristics include (a) if question was answered or not, (b) type of question, namely incidence rates, patient characterization or comparative effectiveness, (c) source of question (email, clinical round, in person communication) and (d) question application (current patient, group of patient or research).

	Original question	Reformatted question	Characteristics	Additional comments from users
1	Does ceftriaxone impacts bilirubin levels in kids under 2 months old?	How many patients under 2 months old got ceftriaxone and what their bilirubin and albumin within a week before and after were? Characterize patients by age, days of ceftriaxone use, gender and conditions (primary diagnosis) since birth.	a) Answered b) Patient characterization and incidence rates c) Clinical rounds d) Current patient	Clinicians reported that the findings aligned with the baseline expectation that ceftriaxone does not cause bilirubin increase. Missing bilirubin measurements after ceftriaxone administration was interpreted as neonates having no symptoms of jaundice. This finding was said to be used in future decision-making.

2	How often Kocuria Marina can be seen in the culture and who are those patients?	How often Kocuria Marina can be seen in the microbal culture? Characterize patients by age, gender, prior disorders, type of culture done (e.g. blood, peritoneal fluid).	a) Answered b) Patient characterization and incidence rates c) Clinical rounds d) Current patient	Clinicians reported that the findings aligned with the baseline expectation that Kocuria Marina did not cause endocarditis and provided more confidence in future clinical decisions.
3	In patients after stenting is there any difference in incidence of in-stent thrombosis after 3, 6, 9 and 12 months of DAPT?	What is the relative risk of in-stent thrombosis within a year after the discontinuation of dual antiplatelet therapy (ticagrelor, prasugrel or clopidogrel in combination with aspirin) in patients	a) Answered b) Comparative effectiveness c) In-person d) Group of patients	Clinicians reported that the findings did not align with the baseline expectation as we observed lower risk of in-stent thrombosis in

		received 3-months DAPT compared to those who received it for 6, 9 and 12 months?		patients receiving 3-months DAPT. As the clinician attributed such a result to the fact that those patients could have been healthier than the others, the findings were unlikely to be used in future decision-making.
4	Patients with a-fib and left atrial appendage closure during cardiac surgery, how many were	What is the relative risk of major cardiovascular events within two years after left atrial appendage closure during cardiac surgery in patients with atrial fibrillation who were on anticoagulants (direct anticoagulants	a) Answered b) Comparative effectiveness c) In-person d) Group of patients	Clinician found that the results obtained on CUMC database were more useful than on the other data sources as we achieved

	anticoagulated (and how did they do)?	or warfarin) after the surgery compared to the patient who discontinued anticoagulants?		balance in studied groups on CUMC. The results aligned with prior believes, are likely to be disseminated and used in future decision-making.
5	What anticoagulants were used for patients with advanced CKD or ESRD (and how did they do)?	What is the relative risk of major cardiovascular events and bleeding within two years after start of anticoagulation therapy in patients with end stage renal disorder treated with warfarin compared to patient treated with apixaban, rivaroxaban or dabigatran?	a) Answered b) Comparative effectiveness c) In-person d) Group of patients	As we could not achieve balance between studied groups, the clinician perceived provided results as inconclusive and would not use them for decision making.

6	How many patients under 18 years old have developed hemolysis within a week after ceftriaxone administration?	How many patients under 18 years old have developed hemolysis (laboratory test or an allergy description of hemolysis) within a week after ceftriaxone administration?	a) Answered b) Patient characterization and incidence rates c) Clinical rounds d) Current patient	Clinician was satisfied with the results but said that capture of hemolysis in EHR might be insufficient. Further studies may be needed to reach final conclusion.
7	Do kids develop allergy to levofloxacin if they are allergic to cephalosporins?	How many patients under 18 years old have a recorded allergic reaction to a cephalosporin (cefazolin, ceftriaxone ceftazidime or cefepime) and levofloxacin within a week? Characterize patients by age, gender, conditions, type of allergy, and	a) Answered b) Patient characterization and incidence rates c) Clinical rounds d) Current patient	Clinician was satisfied with the findings but stated that further study on patient notes may be needed to better capture allergy events.

		concomitant medication use a year prior to the allergic event.		
8	Does metformin impact the risk of atrial fibrillation or other arrythmias in diabetic patients?	What is the relative risk of cardiac arrythmias (supraventricular and ventricular) in patients with diabetes who took different first-line antidiabetic drugs monotherapy (metformin, thiazolidinediones, DPP4 inhibitors, GLP-1 agonists) for at least a year?	a) Answered b) Comparative effectiveness c) In-person d) Research	Clinician was surprised by the high variation in first-line therapy for diabetes was found. Initially the sample size in CUMC was dissatisfying, but an expected number of studied patients was obtained in IBM Marketscan Medicare database.

9	In patients who had baseline indications for anticoagulation and are continued on therapeutic anticoagulation during their stay, do we observe a lesser frequency of thrombotic complications or thromboembolism biomarkers?	What is the relative risk of thrombotic complications (increased d-dimer, acute kidney injury, stroke, acute myocardial infarction, deep venous thrombosis or pulmonary embolism) in patients with baseline indications for anticoagulation on therapeutic anticoagulation regimen during hospitalization compared to those on prophylactic anticoagulation regimen?	a) Answered b) Comparative effectiveness c) Email d) Group of patients	Clinicians considered findings for d-dimer to be more trustworthy compared to findings for deep venous thrombosis due to underreporting of the latter.
10	Do patients with COVID-19 develop	What is the incidence rate of acute kidney injury during hospitalization in patients with no prior end stage renal disorder hospitalized	a) Answered	The findings aligned with clinicians' experience and

	acute kidney injury more often?	with COVID-19 diagnosis? Characterize the patients who developed acute kidney disorder.	b) Patient characterization and incidence rates c) Email d) Research	were said to be likely to be used in practice.
11	Does famotidine have a protective effect for COVID patients?	What is the relative risk of COVID-19 infection in patients on famotidine for at least 3 months prior to COVID test compared to those on proton pomp inhibitors?	a) Answered b) Comparative effectiveness c) Email d) Group of patients	The findings supported the baseline expectation that famotidine is not associated with decreased risk of COVID-19 and therefore provided more confidence in the decision not to prescribe it to prevent COVID-19.

12	How often does Zosyn cause immunosuppression?	What is the incidence rate of immunosuppression (WBC<$4x10^3$/ml) within a week after Zosyn administration in patients with no prior cancer, leukemia, lymphoma and systemic disorders?	a) Answered b) Patient characterization and incidence rates c) Clinical rounds d) Current patient	The findings supported the baseline expectation and are likely to be disseminated and used in future decision-making.
13	How patients with epilepsy and drug resistant epilepsy are treated at CUMC?	Characterize treatment pathways and patient features in patients diagnosed with epilepsy who received anti-epileptic treatment for at least a year at CUMC (at least one exposure every four months).	a) Answered b) Patient characterization and incidence rates c) Email d) Group of patients	Clinician found the pathways patterns aligning with the baseline expectations. Specifically, we found that most of the patients with resistant epilepsy are not recorder as

				such in structured data but can be identified based on their drug exposures.
14	Do different groups of drugs for breast cancer have a protective effect for COVID infection?	What is the relative risk of COVID-19 infection in patients with breast cancer on different types of treatment (estrogen receptor blocker or aromatase inhibitors) prior to COVID test?	a) Answered b) Comparative effectiveness c) Email d) Research	We found no difference in risk of COVID-19 infection between studied groups, but sample size is insufficient to produce reliable estimates.
15	Does glucosamine have a protective effect for COVID infection in patients with osteoarthritis?	What is the relative risk of COVID-19 infection in patients with osteoarthritis on glucosamine compared to patients on diclofenac or oxicams prescribed within 3 months prior to COVID test?	a) Answered Comparative effectiveness c) Email d) Research	We found no difference in risk of COVID-19 infection between studied groups, but glucosamine capture is

				insufficient to produce reliable estimates.
16	Does doxycycline help to prevent COVID?	What is the relative risk of COVID-19 infection in patients with acne on doxycycline compared to those on macrolides?	a) Answered b) Comparative effectiveness c) Email d) Research	We found no difference in risk of COVID-19 infection between studied groups, but sample size is insufficient to produce reliable estimates.
17	What is the risk of transverse myelitis in patients who received influenza vaccine?	What is the relative risk of transverse myelitis and other common vaccine adverse events in patients who received influenza vaccination?	a) Answered b) Comparative effectiveness c) Email d) Research	Preliminary data analysis revealed substantial number of transverse myelitis cases within 6 months after vaccination.

18	What proportion of our patients who get evaluated for kidney transplant end up having a cardiac catheterization and coronary intervention (stenting) before being waitlisted?	What is the incidence rate of cardiac catheterization and stenting in patients evaluated for kidney transplant prior to placing on the waitlist? Characterize patients who got cardiac intervention and who did not by age, gender, comorbidities and time from entering the cohort to an outcome (intervention or placement in the waitlist).	a) Answered b) Patient characterization and incidence rates c) Email d) Group of patients	Clinicians reported that the findings aligned with the baseline expectation that only a small portion of evaluated patients receive coronary intervention.
19	What proportion of our waitlisted patients who get re-tested end up having to undergo cardiac catheterization	What is the incidence rate of cardiac catheterization and stenting in patients placed on waitlist for kidney transplant and received second testing?	a) Answered b) Patient characterization and incidence rates c) Email	Clinicians reported that the findings aligned with the baseline expectation that secondary testing did not

	and coronary intervention based on results of re-testing?		d) Group of patients	result in sufficient number of coronary interventions.
20	How likely are the patients who receive high doses of steroids and insulin for kidney transplant to develop diabetes?	In patients with new kidney transplant who have no prior diabetes, are administered high doses of steroids and require insulin injections during their stay, what is the likelihood of developing incident diabetes within 2 years after insulin administration?	a) Answered b) Patient characterization and incidence rates c) Email d) Group of patients	Initial data exploration revealed that use of steroids increased risk of incident diabetes after renal transplant. Further studies with appropriate comparator groups are needed.
21	Is a combination of tacrolimus and	What is the risk of kidney transplant rejection in patients with new kidney	a) Answered	Clinicians were satisfied with the report and

	belatacept superior to monotherapy in preventing kidney transplant rejection?	transplant on belatacept with tacrolimus compared to those drugs alone within a year after transplant?	b) Patient characterization and incidence rates c) Email d) Group of patients	indicated that they would share the findings and use them in the decision-making.
22	How likely are patients who gain more than 10 pounds after kidney transplant to develop diabetes?	In patients with kidney transplant, is weight gain (more than 10 pounds) associated with new onset of diabetes within two years?	a) Answered b) Patient characterization and incidence rates c) Email d) Group of patients	Clinicians were satisfied with the report, although it did not align with their prior belief.
23	Is dupilumab associated with reduced risk of	What is the relative risk of atrial fibrillation exacerbation in patients with pacemaker on	a) Answered	Initial data exploration revealed no difference in

	atrial fibrillation exacerbation in patients with pacemaker?	duplimab compared to patients who do not receive this drug?	b) Comparative effectiveness c) Email d) Group of patients	atrial fibrillation exacerbation risk, but further studies with a larger sample size are needed.
24	Does measurement of beta hydroxybutyrate level lead to overdiagnosis of diabetic ketoacidosis?	What is the incidence of diabetic ketoacidosis in patients with hydroxybutyrate level measured compared to those who did not receive the test?	a) Answered b) Patient characterization and incidence rates c) Email d) Group of patients	We found increased incidence of diabetic ketoacidosis in diabetic patients after beta hydroxybutyrate blood level measurement was introduced into clinical practice. We have not received the feedback from clinicians yet.

25	In ICU and CCU, does high amount of vasopressors correlates with survival rate?	What is the relative risk of death in patients hospitalized to intensive or critical care unit and prescribed high doses of vasopressors (>30 mcg/min of epinephrine) compared to patients on regular doses of vasopressors?	a) Not answered b) Comparative effectiveness c) In-person d) Group of patients	
26	Are there specific circumstances or patient populations that you prefer to use carbapenems over cefepime for Enterobacter spp?	What are the characteristics of patients with detected Enterobacter (MIC =/<2) who had prior documented *in vitro* susceptibility to cefepime and got prescribed carbapenems?	a) Not answered b) Patient characterization and incidence rates c) Email d) Current patient	
27	How patients with endocarditis and	What were the treatment pathways in patients with endocarditis and negative blood	a) Not answered	

	negative blood culture are treated?	cultures (whenever blood culture was performed prior to the start of antibiotic therapy)?	b) Patient characterization and incidence rates c) Clinical rounds d) Group of patients	
28	How often patients positive for methicillin-sensitive Staphylococcus aureus (MSSA) also get infected with methicillin-resistant Staphylococcus aureus (MRSA)?	How many patients with surveillance culture positive for MSSA (tested through either MRSA screening culture or MSSA/MRSA PCR screen) have any positive culture for MRSA within a year of surveillance culture from any site? Characterize patients by age, gender, prior disorders and type of MRSA culture (either wound, blood etc).	a) Not answered b) Patient characterization and incidence rates c) Clinical rounds d) Group of patients	

29	How often do we see cardiac complications after the initiation of immunotherapy?	How often do we see cardiac complications (diagnosis of myocarditis, troponin > 100 ng/ml, decreased ejection fraction) within 3 months after the initiation of immunotherapy?	a) Not answered b) Patient characterization and incidence rates c) In-person d) Group of patients	

Appendix 5.7 Examples of reports provided by the Data Consult Service

5.7.1 Example of report №1

Original question:

How many patients under 2 months old got ceftriaxone and what their bilirubin and albumin within a week before and after were? Characterize patients by age, days of ceftriaxone use, gender, conditions and procedures since birth.

Our findings:

We found 434 patients under 2 months old with 541 ceftriaxone administrations. 405 patients did not have bilirubin measurements. We observed 11 patients (2.5%) with increased bilirubin: 5 patients who had bilirubin<1.2 mg/dL before ceftriaxone administration and bilirubin>1.2 mg/dL within a week after ceftriaxone administration and 6 patients who had initial abnormal bilirubin measurement and an increased bilirubin within a week after ceftriaxone administration. Other 18 patients had the same or decreased bilirubin measurement.

Table 1. Patient characteristics

		All patients	Patients with increased bilirubin
# of identified patients		434	11
Age		41±13.7 days	30±16.7 days
Gender	Male	254 (58.5%)	7 (64%)
	Female	180 (41.5%)	4 (36%)
Abnormal body temperature or fever		90%	100%
Sepsis		21.4%	63%
Respiratory condition of fetus OR newborn		12.2%	63%

Urinary tract infectious disease	12%	36%
Acute upper respiratory infection	11.5%	9%
Neonatal jaundice	9.9%	36%
Fetal or neonatal effect of maternal infection	8.8%	27%
Prematurity of fetus	6.7%	45%
Viral disease	5.8%	9%
Diarrhea	5%	0%

Figure 2. Distribution of days of ceftriaxone use

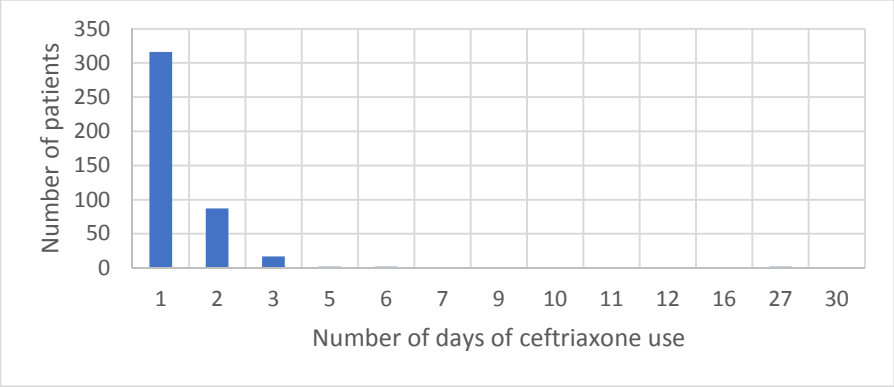


Figure 3. Distribution of bilirubin values, mg/dL before (y-axis) and after (x-axis) an exposure, a dot represents one patient

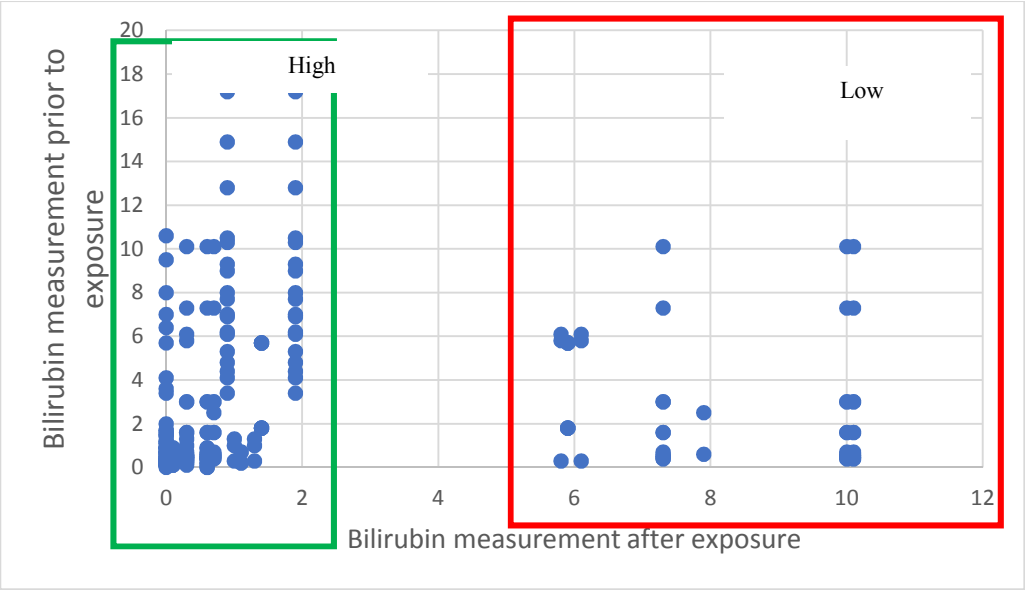
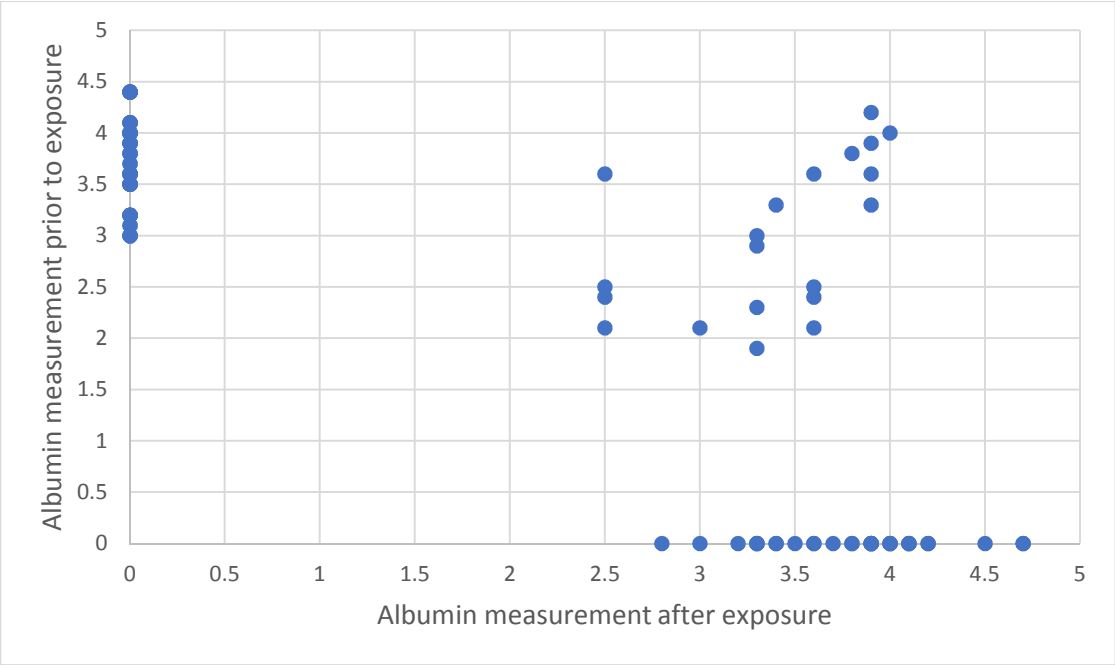


Figure 4. Distribution of maximum albumin values, g/dL before (y-axis) and after (x-axis) an exposure, a dot represents one patient.



Additional Information

1. Concept set definition

1.1. Patients who took Ceftriaxone (RxNorm 2193) and all drugs that contain it.

The list of drugs included following strength:

Ceftriaxone 250 MG Injection – 140 patients

Ceftriaxone 500 MG Injection – 342 patients

Ceftriaxone 1000 MG Injection – 58 patients

1.2. Bilirubin and albumin measurements

Medical Entity Dictionary codes:

4134, 11468, 33673 Direct Bilirubin

5844, 40017 Neonatal bilirubin

27697, 9286, 40945, 36800, 15942, 33694 Total Bilirubin

37546, 36085, 13735, 9820, 27699, 15944 Indirect bilirubin

13153, 13160, 13162, 11303, 36080, 3716, 3829, 4132, 25432, 37541, 17220, 36799, 9821 Albumin serum or plasma

5.7.2. Example of report №2

Original question: Patients with a-fib and left atrial appendage closure during cardiac surgery, how many were anticoagulated (and how did they do)?

Our interpretation: What is the relative risk of major cardiovascular events within two years after left atrial appendage closure during cardiac surgery in patients with atrial fibrillation who were on anticoagulants (DOAC or warfarin) after the surgery compared to the patient who discontinued anticoagulants?

Our findings:

We ran a comparative effectiveness study with propensity score matching on Truven MarketScan Medicaid, Medicare and Columbia University Irving Medical Center data.

We found statistically significant lower risk of stroke within 2 years after left atrial appendage closure during surgery in patients on anticoagulants compared to patients off anticoagulants (odds ratio 0.34, CI 95% 0.14 – 0.88, P-value 0.02) on CUIMC data.

We observed no difference in risk of stroke or MI on neither Medicaid nor Medicaid data, but we did not achieve full balance between cohorts on these two data sources. After propensity score matching, patients were still appreciably different. For example, patients on anticoagulants had fewer cases of bleeding before the surgery, fewer cases of respiratory failure, pneumonia or other respiratory disorders, chronic liver disease (supplementary materials 3.3).

For reference, an unadjusted comparison (Tables 1 and 2) showed that anticoagulants were more frequently used in patients with rheumatic heart disease, mitral valve disorders, combined valve disorders, pulmonary hypertension, but these results are subject to confounding bias and are less trustworthy than the adjusted ones. Overall, anticoagulants were used less frequently when patients had

bypass surgery and more frequently when the surgery was related to valve procedures. Anticoagulants were used more frequently in cases of any re-operation. We also observed that history of stroke did not seem to influence anticoagulants prescription.

Methods:

This was a retrospective cohort study using Cox proportional hazard model to assess the hazard ratios (HRs) between the target cohort (patients on anticoagulants) and comparator cohorts (patients off anticoagulants). We used clinical concepts from Supplementary materials №1 to identify patients based on the cohort definitions in Supplementary materials №2. Once this base cohort was established, patients were evaluated to determine if they had an outcome of interest (AMI, stroke or VTE) during a 2-year follow-up period. Adjustment for baseline confounders was performed by fitting a propensity model and using the resultant propensity scores to match the target and comparator cohorts using variable ratio matching (Supplementary materials №3). To identify potential residual bias in HR estimates, we used negative control outcomes (i.e., not believed to be caused by anticoagulants). HRs were computed for these negative controls and used to compute calibrated p□values for outcomes.

We also characterized patients in the target and comparator cohorts before propensity score matching (Table 1 and Table 2).

Table 1. Type of surgery associated with left atrial appendage closure, on anticoagulants/off anticoagulants, red – higher in the group on anticoagulants, green – higher in group off anticoagulants.

Type of surgery	**Medicaid data**	**Medicare data**	**CUIMC data**
Coronary Artery Bypass, 1 artery	21/34%	29/37%	5/23%
Coronary Artery Bypass, 2 arteries	7/16%	12/13%	6/6%
Coronary Artery Bypass, 3 arteries	5/11%	6/6%	2/4%
Replacement, mitral valve	45/20%	22/14%	28/26%
Replacement, aortic valve	19/11%	32/29%	26/36%
Repair atrial septal defect, secundum	8/3%	5/2%	7/6%

Reoperation, bypass or valve procedure	7/2%	3.2/2.4%	7/3%

Table 2. Patient characteristics (unadjusted), Medicaid/Medicare/CUIMC data; green represents relatively low values, red – relatively high. Colors **do not** imply statistical significance and we did not test if the difference between groups is significant.

		On anticoagulants	**Off anticoagulants**
# of identified patients		122/389/350	277/337/91
% of patients who experienced stroke within 2 years		11/7/11.2%	10/5/15.6%
Mean age		$60 \pm 11/74 \pm 6/67 \pm 12$	$60 \pm 13/75 \pm 6/68 \pm 11$
Race	Black	21/-/2%	22/-/2%
	White	52/-/53%	60/-/50%
	Other	27/-/45%	18/-/48%
Ethnicity	Hispanic or Latino	2/-/37%	1/-/13%
	Not Hispanic or Latino	98/-/45%	99/-/47%

Gender	Male	49/58/60%	53/69/60%
	Female	52/42/40%	47/31/40%
Comorbidities	Hyperlipidemia	72/88/41%	73/92/53%
	Anemia	74/71/13%	74/73/14%
	Heart failure	75/64/46%	78/65/44%
	Peripheral vascular disease	60/77/49%	60/78/64%
	Pulmonary hypertension	51/32/31%	31/27/26%
	Rheumatic heart disease	68/50/28%	44/43/23%
	Diabetes mellitus II	40/42/21%	50/47/27%
Non-rheumatic mitral valve disease		75/68/52%	56/56/35%
Rheumatic disease of mitral valve		49/33/22%	30/24/15%
Non-rheumatic tricuspid valve disease		45/33/14%	27/33/10%
Rheumatic disease of tricuspid valve		25/15/8%	14/13/5%
Disorders of both mitral and tricuspid valves		40/22/-%	19/22/-%

Combined disorders of mitral, aortic and tricuspid valves	14/14/-%	8/11/-%
H/o stroke	14/11/7%	17/14/9%
CHADS2VASc	3.5 ±1.6/4.7±1.6/4±1.3	4 ±1.7/4.9±1.6/4.2±1.5

Supplementary materials

1 Concept set definition

1.1 Anticoagulants

All drugs that contain apixaban, dabigatran, rivaroxaban or warfarin.

1.2 For VTE, AMI and stroke we used definitions from the LEGEND hypertension study (published here: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)32317-7/fulltext#sec1)

1.3 For left atrial appendage closure, we used the following codes with their descendants:

02L70ZK Occlusion of Left Atrial Appendage, Open Approach

02L70DK Occlusion of Left Atrial Appendage with Intraluminal Device, Open Approach

02L70CK Occlusion of Left Atrial Appendage with Extraluminal Device, Open Approach

02B70ZK Excision of Left Atrial Appendage, Open Approach

02570ZK Destruction of Left Atrial Appendage, Open Approach

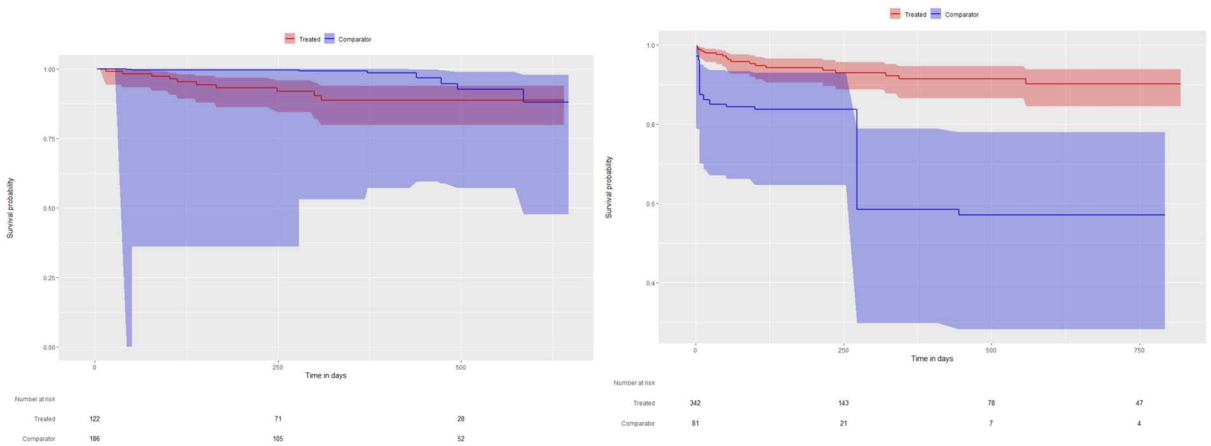
2 Cohort definition

We defined patients with atrial fibrillation on anticoagulant therapy after LAAC during surgery as those who had had a diagnosis of atrial fibrillation prior to LAAC and were administered anticoagulants within 7 days after a LAAC procedure. The latter should occur on the same day with any cardiac surgery.

We defined patients with stroke, AMI or VTE as those who had corresponding codes associated with an inpatient or emergency room visit.

3 Additional study-related information

3.1 Kaplan-Meier survival plots



.

Figure 1. KMP showing the probability of stroke on Medicaid data

Figure 2. KMP showing the probability of stroke CUIMC data

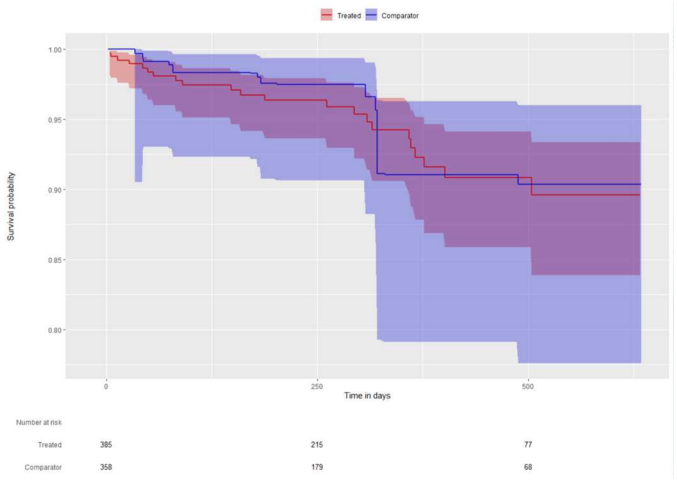


Figure 3. KMP showing the probability of stroke on Medicare data .

3.2 Propensity score matching

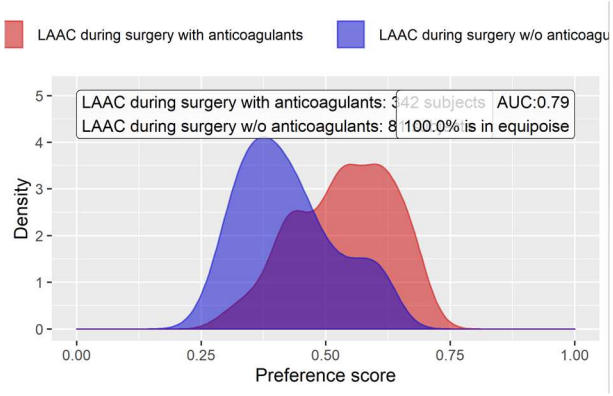


Figure 5. PS matching on CUIMC data

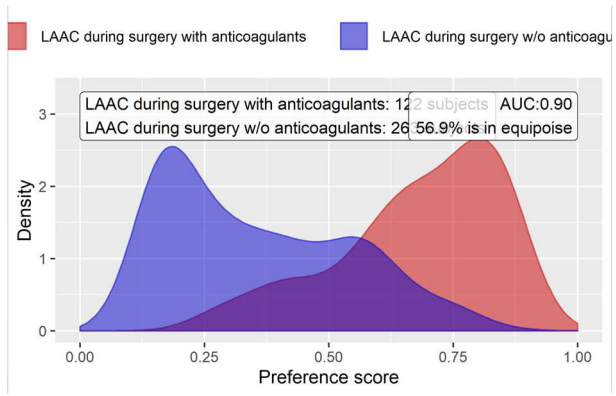


Figure 6. PS matching on Medicaid data

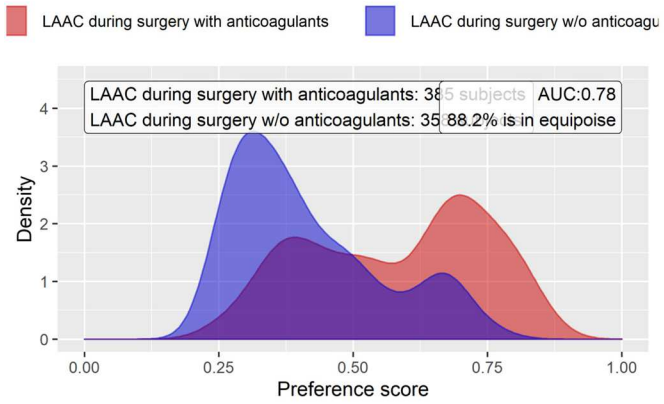


Figure 7. PS matching on Medicare data

Percent of patients in equipoise indicates the percent of patients who had the preference score (probability of receiving target treatment) within 0.25 – 0.75 range. We estimated patients' propensity scores based on their drug, condition, procedure and measurement records as well as CHA2DS2-VASc and Charlson comorbidity index, **assuming that this percent of patients from the target and comparator groups are comparable**. We identified more than 10,000 covariates for each analysis.

Top covariates for which we did not achieve good balance between the target and the comparator groups include: respiratory failure and insufficiency, pneumonia and other respiratory disorders, history of bleeding and chronic liver disease.

3.3 Covariate balance before and after propensity score matching.

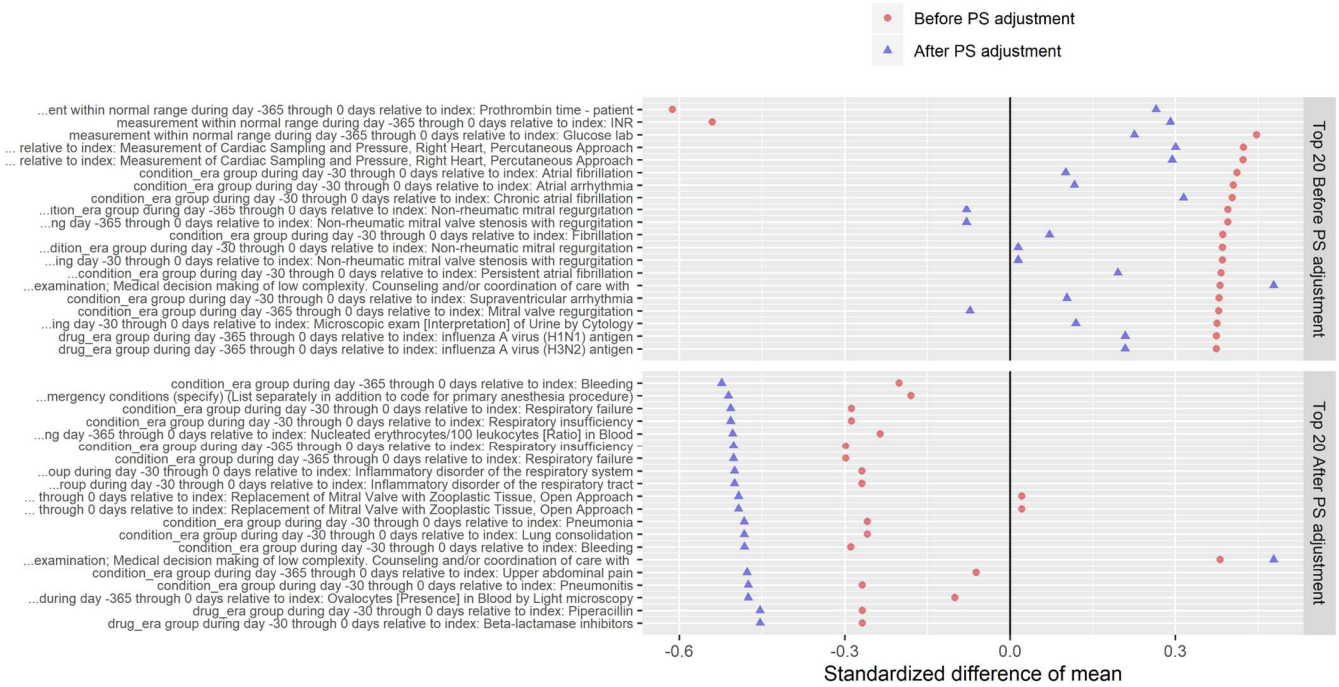


Figure 8. Covariate balance for stroke, CUIMC data

Figure 9. Covariate balance for stroke, Medicaid data

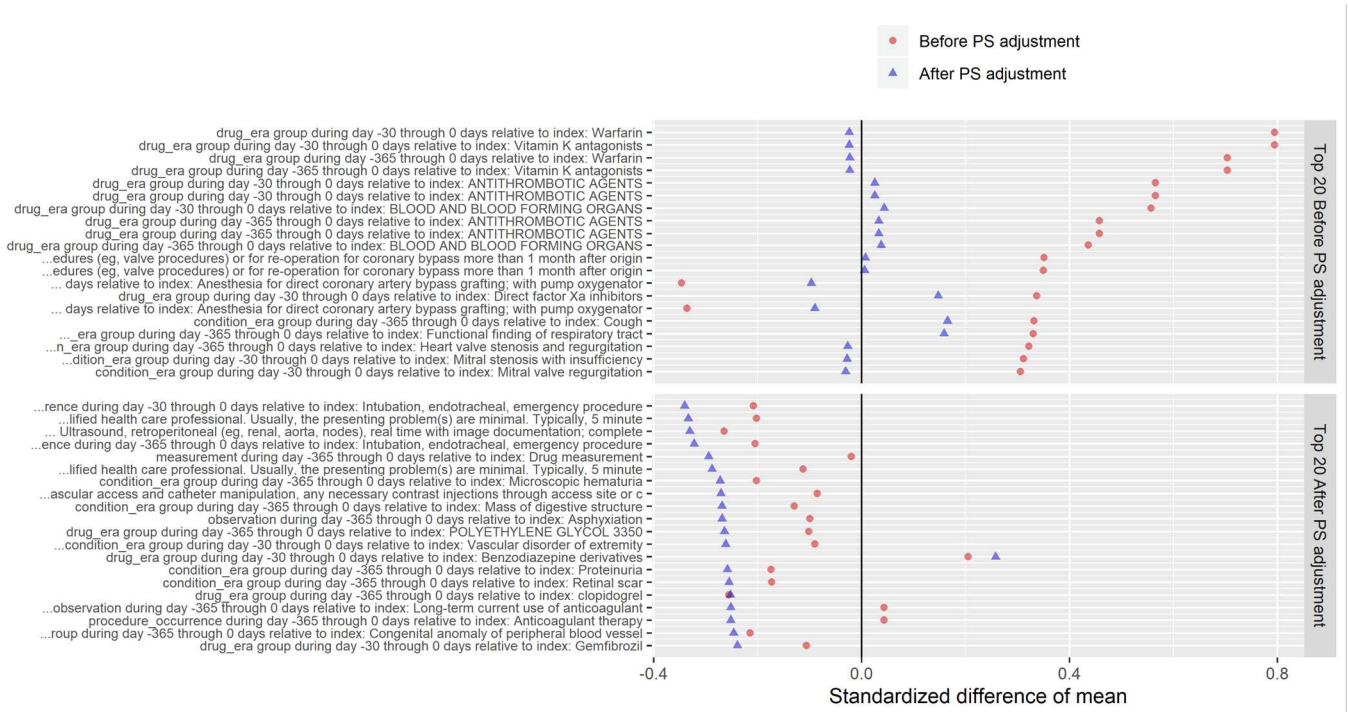


Figure 10. Covariate balance for stroke, Medicare data