25

Distributed Networks of Databases Analyzed Using Common Protocols and/or Common Data Models

Sengwee Toh[1], Nicole Pratt[2], Olaf Klungel[3], Joshua J. Gagne[4], and Robert W. Platt[5]

[1] *Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA*
[2] *Quality Use of Medicines and Pharmacy Research Centre, School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, South Australia, Australia*
[3] *Division of Pharmacoepidemiology & Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, The Netherlands*
[4] *Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA*
[5] *Departments of Epidemiology, Biostatistics, and Occupational Health, and of Pediatrics, McGill University, Montreal, Quebec, Canada*

It is now common to analyze large and complex electronic healthcare databases – created as part of regular business operations or routine clinical care – to assess the safety and effectiveness of medical products. Pharmacoepidemiologic studies have traditionally analyzed information available in single databases. However, single-database studies may not be sufficient to answer certain clinical questions, especially when the exposure or outcome is rare, when the goal is to study the treatment effect in specific subgroups, or when the objective is to identify a sufficiently large number of exposed patients within a relatively short time window (e.g., during the early months following the approval of a new medical product).

Thanks to the increase in the number of data sources, and the improvement in the quality of and ease of access to these data sources, multidatabase studies are now feasible and ubiquitous in pharmacoepidemiologic research. There are several ways to conduct multidatabase studies. An intuitive approach is to pool the databases or the derived analytic datasets centrally for analysis. However, centralized pooling of databases that contain detailed individual-level data is not always possible for several reasons, including concerns about patient privacy, data security, unauthorized uses of data, and potential disclosures of sensitive institutional or business information. A distributed approach, in which databases are not combined centrally but rather stored in different physical locations under the direct control of the participating sites, is becoming increasingly preferred.

In this chapter, we describe the design, development, implementation, strengths, and challenges of distributed data networks (DDNs). We begin with a brief description of select DDNs in pharmacoepidemiology. We then discuss the types of research questions that DDNs are designed to address. We examine the methodologic and data issues unique to DDNs, and progress that has been accomplished to address these issues. We conclude with a discussion about some of the future directions for DDNs.

Here we first define three key terms that are central to the chapter:

- *Distributed data network*: two or more data sources stored in different physical locations

Pharmacoepidemiology, Sixth Edition. Edited by Brian L. Strom, Stephen E. Kimmel and Sean Hennessy.

under the direct control of the participating data partners.
- *Common data model (CDM)*: a data model that generally includes a set of standardized data files and variables, adopted by all data sources participating in a DDN.
- *Common protocol*: a study protocol that typically includes detailed description of key design and analytic parameters of the study, implemented by all data sources participating in a DDN study.

Examples of Distributed Data Networks in Pharmacoepidemiology

DDNs have been in existence for more than 20 years. Drawing on more than two decades of experience, the DDNs today are more sustainable, efficient, and diverse. This section briefly describes a number of DDNs designed to conduct pharmacoepidemiologic research, medical product safety surveillance, or comparative effectiveness research using electronic health data collected as part of routine healthcare delivery. Table 25.1 provides a summary of their key characteristics.

Asian Pharmacoepidemiology Network (AsPEN)

Established in 2008, AsPEN is a multinational research network formed to provide a mechanism to support the conduct of pharmacoepidemiologic research and to facilitate more rapid identification and validation of emerging safety issues among the Asia-Pacific countries [1,8]. AsPEN is a collaboration of 8 countries involving 12 databases formalized as a Special Interest Group of the International Society for Pharmacoepidemiology. It has piloted a number of approaches to conduct distributed studies including a common protocol, a standard analytic program [9–12], and translation to a CDM developed by the Observational Medical Outcomes Partnership (OMOP; see later) [13].

Canadian Network for Observational Drug Effect Studies (CNODES)

CNODES is a distributed network of Canadian research teams designed to provide evidence to Canadian stakeholders, in particular Health Canada, on drug safety in the Canadian context [2]. It is one of four collaborating centers supported by the Drug Safety and Effectiveness Network (DSEN) of the Canadian Institutes of Health Research. Queries from stakeholders are prioritized by DSEN head office and sent to the CNODES coordinating center. CNODES teams conduct analyses of a distributed network of Canadian and international databases, and report both to stakeholders and via published literature.

Health Care Systems Research Network (HCSRN)

Established in 1994, HCSRN (formerly known as the Health Maintenance Organization Research Network, HMORN) is a consortium of 18 integrated delivery systems and health plans designed to facilitate multidatabase collaborative research [3]. Compared to other DDNs, HCSRN is unique because it is not created to address specific clinical or research questions. Instead, it is "multipurpose" and supports a wide range of research and surveillance activities. HCSRN is the foundation on which several large-scale collaborative projects are built and maintained, including the Vaccine Safety Datalink (VSD; more later) [14], the Cancer Research Network [15], the Mental Health Research Network [16], and the Cardiovascular Research Network [17]. HCSRN is often considered one of the best examples of a sustained DDN [3]. Its distributed network architecture and CDM (known as the Virtual Data

Table 25.1 Select examples of distributed data networks in pharmacoepidemiology.

	AsPEN [1]	CNODES [2]	HCSRN [3]	PCORnet [4]	PROTECT [5]	Sentinel [6]	VSD [7]
Number of data partners	12	9	18	>80	14	18	9
Total population	220 million	35 million (Canada)	16 million	100 million	100 million	293 million	9 million
Type of data	Claims	Claims, EHRs	Claims, EHRs	Claims, EHRs	Claims, EHRs	Claims, EHRs	EHRs
Geography	Asia-Pacific	Canada, US, and UK	US and Israel	US	European Union	US	US
Funding source	None	CIHR	Various	PCORI	IMI	FDA	CDC
Primary mission	Medication safety research	Medical product safety research	Multipurpose	Patient-centered outcomes research	Medication safety research	Medical product safety surveillance	Vaccine safety surveillance
Common data model	Yes	Yes	Yes	Yes	No	Yes	Yes
Common study protocol	No	Yes	Yes	Yes	Yes	Yes	Yes
Common statistical analysis plan	Yes	Yes	Yes	Yes	Yes	Yes	Yes

AsPEN, Asian Pharmacoepidemiology Network; CDC, Centers for Disease Control and Prevention; CIHR, Canadian Institutes of Health Research; CNODES, Canadian Network for Observational Drug Effect Studies; EHRs, electronic health records; FDA, Food and Drug Administration; HCSRN, Health Care Systems Research Network; IMI, Innovative Medicines Initiative; PCORI, Patient-Centered Outcomes Research Institute; PCORnet, National Patient-Centered Clinical Research Network; PROTECT, Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium; VSD, Vaccine Safety Datalink.

Warehouse) [18] have been adopted by other DDNs, such as the Sentinel system (see later).

National Patient-Centered Clinical Research Network (PCORnet)

Launched in 2013, PCORnet is a network of networks that includes 13 Clinical Data Research Networks, 20 People-Powered Research Networks, 2 Health Plan Research Networks, and a coordinating center in the US [4]. The Clinical Data Research Networks are primarily comprised of healthcare delivery systems, the Health Plan Research Networks are shepherded by two national insurers, and the People-Powered Research Networks are mainly led by patient and caregiver organizations. PCORnet is designed to support both randomized trials and observational studies. It currently includes electronic health record (EHR) or administrative claims data from more than 100 million individuals and has access to over 40 million patients who could be recruited into pragmatic clinical trials.

Pharmacoepidemiologic Research on Outcomes of Therapeutics by a European Consortium (PROTECT)

Initiated in 2009 and ended in 2015, PROTECT was a joint undertaking by the European Union (EU) and pharmaceutical industry as part of the Innovative Medicines Initiative [5]. Its 35 partners, including academics, regulators, small and medium enterprises, and member companies of the European Federation of Pharmaceuticals Industries and Associations, were coordinated by the European Medicines Agency. The overall objective of PROTECT was to address limitations of current methods in the field of pharmacoepidemiology and pharmacovigilance. As part of this work, a network of electronic healthcare databases was established to conduct multicountry, multidatabase, drug safety studies. Currently, several former public partners of PROTECT are continuing their collaboration with additional public partners in the European Research Network for Pharmacoepidemiology and Pharmacovigilance, allowing access to a broad variety of datasets (general practice, hospital pharmacy/laboratory, pharmacy, hospitalization, claims, questionnaires, and biological samples) covering six EU countries (Spain, UK, Italy, the Netherlands, Denmark, and France) and records from approximately 100 million active patients to address various research questions.

Sentinel System

The Sentinel system is funded by the US Food and Drug Administration (FDA) as a national medical product surveillance system mandated by the US Congress in the FDA Amendments Act of 2007 [6,19,20]. Initiated as a pilot program called Mini-Sentinel in 2009 [21], the system includes a distributed network of 18 data partners that provide access to administrative claims and EHR information from over 290 million cumulative patient identifiers. A feature of Sentinel is its ability to conduct rapid descriptive and inferential analysis using preprogrammed, pretested, and customizable analytic tools [22–25]. These analytic tools and other Sentinel-related materials (e.g., protocols, reports, data model) are all available in the public domain.

Vaccine Safety Datalink (VSD)

Funded by the US Centers for Disease Control and Prevention since 1990, VSD monitors the safety of vaccines using EHR databases from a network of nine delivery systems and health plans [7,14,26,27]. While VSD initially used a centralized data model in which the data partners submitted de-identified analytic datasets for centralized analysis, it switched to a more sustainable DDN model in 2001 [28]. A unique feature of VSD is its ability to provide near

real-time surveillance of vaccine safety. Researchers apply study design and statistical methods appropriate for sequential surveillance to analyze weekly updated data. The analysis takes into account data lag and incompleteness due to frequent refreshes of the data [29,30].

Others

OMOP was a partnership between the FDA and the Pharmaceutical Research and Manufacturers of America established to inform the appropriate use of electronic healthcare databases for studying the effects of medical products [13]. In achieving that goal, OMOP initiated a series of experiments to investigate the feasibility and validity of conducting fully automated assessments of medical product safety [31,32]. The initiative also created a CDM and a suite of CDM-compatible analytic tools. The OMOP experiment was completed in 2013, but much of its collaborative work continues within Observational Health Data Sciences and Informatics (OHDSI) [33].

The "Exploring and Understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge" (EU-ADR) project was launched in 2008 with the aim to leverage information from various EHR databases in Europe to produce a computerized integrated system for the early detection of drug safety signals [34]. The same approach and some of the same databases were used in study-specific networks, including the Safety of Nonsteroidal Anti-inflammatory Drugs (SOS) [35], Arrhythmogenic Potential of Drugs (ARITMO) [26], Safety Evaluation of Adverse Reactions in Diabetes (SAFEGUARD) [37], Global Research in Paediatrics (GRIP) [38], and Accelerated Development of Vaccine Benefit-risk Collaboration in Europe (ADVANCE) [39] projects. Apart from EU-ADR, these collaboratives have included several of the same EU datasets and developed study-specific CDMs.

Clinical Problems to Be Addressed by Pharmacoepidemiologic Research

As with multidatabase studies that analyze centrally pooled information, DDN studies support analyses that cannot typically be done with one data source. Examples include assessments of rare exposures, rare outcomes, treatment effect heterogeneity in specific subpopulations, and surveillance of newly approved medical products (Table 25.2).

Assessment of Rare Exposures

Examples of rare exposures include, but are not limited to, drugs used to treat orphan diseases, defined as conditions that affect fewer than 200 000 individuals (US definition) or 5 in 10 000 individuals (EU definition). DDNs allow studies of drugs indicated for orphan diseases [59,60]. For example, several People-Powered Research Networks within PCORnet, such as the Phelan-McDermid Syndrome Data Network [61], are leveraging patient-generated information and the electronic health data within Clinical Data Research Networks to generate evidence about disease progression and treatments for these conditions.

Assessment of Rare Outcomes

An example of rare outcomes is Guillain-Barré syndrome, which occurs in 1–2 per 100 000 person-years [62]. An adequately powered study to examine the association between a vaccine and Guillain-Barré syndrome requires information from millions of individuals from multiple databases. For example, to assess the risk of Guillain-Barré syndrome following receipt of a quadrivalent human papillomavirus vaccine, a VSD study used six databases to identify males and females aged 9–26 years who received the vaccine from 2006 to 2015 [58]. One confirmed case of Guillain-Barré syndrome within 42 days

Table 25.2 Select examples of studies conducted within distributed data networks.

Network	Select studies
AsPEN	• Antipsychotic use and risk of acute hyperglycemia [9] • Thiazolidinedione use and risk of heart failure across ethnic groups [11] • Cardiac safety of methylphenidate among pediatric patients with ADHD [40]
CNODES	• Statin use and risk of acute kidney injury [41] • Incretin-based drug use and risk of heart failure [42] • Occurrence of pregnancy during isotretinoin therapy [43]
HCSRN	• Lipid-lowering drug use and risk of rhabdomyolysis [44] • Prenatal antidepressant exposure and risks of congenital malformations [45] • ADHD medication exposure and risk of serious cardiovascular events [46]
PCORnet	• Aspirin dosing and secondary prevention of atherosclerotic cardiovascular disease [47] • Antibiotic use and weight outcomes in children [48] • Long-term benefits and risks of bariatric procedures [49]
PROTECT	• Antibiotic use and risk of acute liver injury [50] • Antidepressant use and risk of hip fracture [51] • Antiepileptic drug use and risk of suicidality [52]
Sentinel	• Antihyperglycemic use and risk of acute myocardial infarction [53] • Dabigatran use and risks of bleeding and cardiovascular events [54] • Rotavirus vaccination and risk of intussusception [55]
VSD	• Thimerosal exposure and risks of neuropsychological outcomes [56] • Safety of H1N1 and seasonal influenza vaccines [57] • Quadrivalent human papillomavirus vaccination and risk of Guillain-Barré syndrome [58]

ADHD, Attention-deficit hyperactivity disorder; AsPEN, Asian Pharmacoepidemiology Network; CNODES, Canadian Network for Observational Drug Effect Studies; HCSRN, Health Care Systems Research Network; PCORnet, National Patient-Centered Clinical Research Network; PROTECT, Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium; VSD, Vaccine Safety Datalink.

following vaccination was confirmed among over 2.7 million vaccinees.

Assessment of Treatment Effect Heterogeneity

Certain treatments may have different effectiveness or safety profiles in patients with specific characteristics. Combining information from multiple databases allows researchers to have enough sample sizes in subsets of populations (e.g., children, the elderly, individuals with a history of heart failure). Results from studies of databases that cover demographically and geographically diverse populations also provide better generalizability. For example, AsPEN conducted a study to determine whether the risk of edema or heart failure associated with thiazolidinediones was different between Caucasian and Asian populations, potentially due to differences in the prevalence of metabolizing enzymes in these ethnic groups [11]. In another example, a Sentinel study examined the associations between antihyperglycemic treatments and risk of hospitalized heart failure among diabetes patients with and without a history of cardiovascular disease [63]. The European Research Network for Pharmacoepidemiology and Pharmacovigilance

is currently investigating the risk of major bleeding associated with direct oral anticoagulants in targeted clinical and demographic subgroups for which variations in plasma concentrations might affect the safety of the products [64].

Postmarket Surveillance of Newly Approved Medical Products

An increasingly common scenario that warrants the use of multiple databases is postmarket surveillance of the safety of newly approved medical products. The goal is to monitor new medical products as postmarket experiences of their use accrue in routine clinical practice. The number of users in a single database is usually low in the early postapproval phase, so multiple databases are required to support an informative analysis. Prospective, sequential analysis of cumulating data can be addressed by appropriate statistical techniques [65–67]. For example, the Sentinel system has leveraged its DDN to complete prospective postmarket surveillance of two newly approved medical products and select health outcomes. One assessed the association between saxagliptin (an oral antihyperglycemic agent) and acute myocardial infarction following the approval of the drug in 2009 [53]. The other examined the associations between rivaroxaban and ischemic stroke, intracranial hemorrhage, and major gastrointestinal bleeding after the oral anticoagulation drug was approved in 2011 [68].

Methodologic Problems to Be Addressed by Pharmacoepidemiologic Research

Pooling of individual-level data, often in a format that is stripped of direct identifiers, had traditionally been the default approach used in multidatabase analyses. Sharing of de-identified individual-level data is generally feasible in the presence of proper governance, appropriate data use or sharing agreements, and established collaborative relationships [69]. However, concerns about patient privacy and confidentiality, unauthorized uses of transferred data, and unintended disclosures of sensitive corporate or institutional information have made data sharing increasingly more challenging in practice, specifically in newly formed collaborations or projects that perform a large number of studies [70–72]. Contractual agreements between health plans and some of their members may further restrict sharing of individual-level information with other entities for secondary purposes such as research.

A DDN architecture addresses some of the concerns associated with pooling of individual-level data [70–72]. A typical DDN generally has the following features:

- There is one or more coordinating center(s).
- Data partners maintain physical control of their data.
- Data partners have the ability to review and approve each data request.
- Data partners have the ability to review the output before sharing it with the requester.
- Data partners can opt out of any data request at any time.

These features offer data partners more autonomy in multidatabase studies. They allow data partners to evaluate their ability or willingness to share their data with the requester at various steps of the request. More importantly, a DDN approach keeps the data close to the individuals who know the data best. The data partners can advise on the appropriate use of the data and help investigate data anomalies and interpret findings.

Challenges in Distributed Data Networks

DDNs also come with challenges, some of which are common across all multidatabase studies and others unique to the distributed environment.

There could be heterogeneity in data quality, data completeness, coding system, and patient population. Neither a centralized nor a distributed system is immune to these issues, but they can be more difficult to identify or diagnose in a distributed environment. Compared to single-database studies, multidatabase studies often involve additional administrative and governance issues, such as the need for multicenter ethics review and data-sharing agreements. These issues may sometimes (but not always) be more resource intensive in DDNs than a centralized system, depending on the type of analysis and information shared. There may be a need for more frequent communications between the coordinating center and participating sites in DDNs, which can create delays. In multinational DDNs, the participating data sources may have different languages, different availability of medical products, wider variation in clinical practice and drug utilization, and different responses to medical products due to ethnic or genetic variations. The conduct of the statistical analysis is generally more complicated in DDNs, because individuals responsible for the analysis do not have full access to the source data from all the participating sites. Although DDNs increase overall sample size, the larger sample size does not necessarily help improve control for confounding, since confounding control typically occurs separately within each site.

Currently Available Solutions

To facilitate the conduct of studies, existing DDNs organize themselves differently based on their resources, needs, expertise, and data infrastructure. On one end of the spectrum are DDNs that employ a common protocol and a CDM approach. At the other end of the spectrum are DDNs that have neither a common protocol nor a CDM. There are also DDNs that adopt a common protocol approach without a CDM. These options lie on a continuum and do not represent all the possible scenarios. For example, a DDN can develop a CDM for some of its data partners but not the others. Each of these options has its unique strengths and limitations (Table 25.3). However, some offer clear advantages over the others in many scenarios. In particular, a DDN that has neither a CDM nor a common protocol approach is typically less efficient than the other systems. Table 25.1 summarizes the approaches employed by some of the DDNs.

Common Data Model, with or without a Common Protocol

Some DDNs have all participating data partners convert their source data into standardized data formats, often known as a CDM. The CDM specifies a uniform data file structure and data element naming conventions and definitions across all databases. There are several CDMs in use, including for Sentinel [72], PCORnet [73], HCSRN [18], and OMOP [74]. The first versions of the Sentinel and OMOP CDMs were modeled in part on the HCSRN CDM, and the Sentinel CDM served as the backbone of the PCORnet CDM. CNODES has implemented the Sentinel CDM in four databases, is working to implement it network-wide, and has initiated queries using the CDM.

There is a general misconception that a CDM is a "lowest common denominator" approach, which reduces the data elements in a DDN to only variables common across all databases. In reality, data partners with more information can populate additional tables or variables for use in specific studies. For example, both the HCSRN and Sentinel CDMs allow data partners with clinical information from EHRs to populate additional tables on vital signs and laboratory test results [18,72].

However, certain information may be lost during the standardization process. This may occur when the data elements are available in multiple coding systems and researchers attempt to map across these systems. For example, US databases primarily use the International

Table 25.3 Strengths and limitations of various structures of distributed data networks.

Common data model	**Yes**	**Yes**	**No**	**No**
Common protocol	**Yes**	**No**	**Yes**	**No**
Upfront data infrastructure investment	Substantial	Substantial	Minimal	Minimal
Site-specific statistical programming effort	Minimal to moderate	Minimal to moderate	Moderate to substantial	Moderate to substantial
Ability to develop preparameterized, reusable tools	Yes	Yes	Limited	Limited
Ability to assess database heterogeneity	Yes	Yes	Yes	Yes
Ability to perform analysis tailored to individual databases	3 or 4 (worst)	2 or 3	2 or 3	1 (best)
Ability to ensure consistent analysis across databases	1 (best)	2 or 3	2 or 3	4 (worst)
Study-specific data management and cleaning	Minimal to moderate	Minimal to moderate	Substantial	Substantial
Speed of study-specific analysis	1 (fastest)	2 or 3	2 or 3	4 (slowest)
Marginal cost per study	1 (lowest)	2 or 3	2 or 3	4 (highest)
Reproducibility/validation across sites	1 (best)	3	2	4 (worst)

Classification of Diseases, 9th or 10th Revision, Clinical Modification (ICD-9-CM or ICD-10-CM) coding systems to record diagnoses, while the general practice databases in the UK document diagnoses using Read codes. It is possible to map and standardize these coding systems, but doing so may lead to some information loss or misclassification. In the presence of multiple coding systems, it is still possible to develop a CDM while preserving the fidelity or granularity of the source information for use in actual studies. In the diagnosis example earlier, the CDM will have a variable that contains the specific diagnosis codes (G30z.00 or 410.00) and an additional variable that indicates the code type (Read or ICD-9-CM). DDN studies that analyze against the CDM can then use the two variables together to define the study parameters. In the presence of multiple coding systems, input from researchers and others familiar with the data is required, either during the mapping process or when conducting a study.

DDNs with a CDM almost always conduct their studies with a common protocol (more later). Analyzing CDM-backed databases with a common protocol allows study-specific data checking, management, and analysis to be done via identical computer programs that can be developed and beta-tested by a smaller group of individuals. This helps reduce programming burden at sites, minimizes opportunities for errors across participating sites, and ensures consistent analysis across databases. However, the centrally developed computer programs have to accommodate differences in computing environments (e.g., different operating systems, software versions) to allow successful execution across all participating sites. In addition, a coding error that occurs in a centrally developed program will have an impact on all sites.

On the rare occasions that a common protocol is not developed in the presence of a CDM, the data partners presumably would have more flexibility in answering the study question. This would allow certain data partners that have more data elements in the CDM to include them in their analysis. However, it is worth noting that a common protocol can be developed in a way that also allows database-specific analysis, for example through a semi data-adaptive approach like high-dimensional propensity score analysis, in which the propensity score is built individually in each database using available information rather than using a common set of variables [23].

Common Protocol, with or without a Common Data Model

In a common protocol approach, a protocol is developed, often collaboratively among participating sites, for implementation across the DDN. As already discussed, employing a common protocol approach in the presence of a CDM generally allows the study to be conducted more efficiently, as programming burden is limited to one site rather than having each site develop *de novo* code. In the absence of a CDM, the common protocol is generally less prescriptive, to allow data partners to define the measurement of exposure, outcome, covariates, and other study parameters based on the information available in their databases. For example, in a study of rivaroxaban, the protocol will specify the exposure of interest, but each data source will identify rivaroxaban exposure based on its coding system (e.g., National Drug Codes or Anatomical Therapeutic Chemical Classification System).

As individuals who are most familiar with the data are actively involved in the implementation of the protocol, the study can accommodate the differences in coding practices, data quality and completeness, and other idiosyncratic issues associated with each database. The disadvantage of this approach is that it can lead to variations in the interpretation of the protocol, which may artificially inflate the heterogeneity across sites or affect the robustness of results. In addition, each site is required to have adequate

programming resources to conduct its own analysis. Coordination across the DDN during the study can be intensive in order to resolve any discrepancies and to ensure consistent interpretation of the analysis plan. However, this can be accomplished by using a detailed statistical analysis plan and a phased analysis, in which, for example, analyses are reviewed at various stages in the process (e.g., after baseline tables are populated, after propensity scores are estimated). Protocol refinements and modifications, as well as site-specific amendments (should initial analyses uncover differences in prescription patterns across sites), can help eliminate or explain discrepancies between sites. Some DDNs employ a blinding procedure to mask the results from participating data sources to facilitate more objective assessment of heterogeneity across sites. These processes to improve consistency require substantial time investment by analytic personnel. The common protocol approach can be particularly onerous for DDNs that do not share a common language or coding system across data partners.

The Necessity of Having a Common Data Model

Creating a CDM is a substantial undertaking. It requires considerable upfront investment on data infrastructure, in particular the extraction, transformation, and loading of the source data to a CDM. Additionally, the ongoing maintenance of the CDM can be burdensome, particularly as new versions of the CDM are needed, which can occur when there are new analytic requirements or changes in certain data elements. Each site must also routinely convert its new data into the CDM. It is generally easier to create a CDM for databases that contain the same type of information and coding system (e.g., claims data coded in the ICD-10-CM system). Developing a CDM for disparate data sources (e.g., claims databases and EHR databases) or databases with different coding systems is more challenging, but is possible through the use of mapping algorithms. However, as discussed earlier, there may be information loss or misclassification if mapping is required.

In general, it will be worthwhile to develop a CDM if the DDN is designed to conduct multiple studies. It may also be useful to convert the source data into a specific CDM to leverage available software or tools that are compatible with the CDM. For example, Sentinel, PCORnet, and OHDSI have developed a suite of analytic tools that can be executed within databases that use their CDMs. From the scientific validity perspective, the amount of data management, quality assurance, and harmonization for a given multidatabase study is similar regardless of the data network architecture. The CDM approach spends more resources upfront on data harmonization and quality assurance, so that downstream studies can be done more efficiently. However, the cost of establishing and maintaining the CDM is the same for 1 study as it is for 100 studies. The CDM achieves an economy of scale when the number of studies supported by the CDM is sufficiently large. In principle, the marginal cost of conducting a study is lower in DDNs with a CDM than in DDNs without a CDM when the number of studies is large.

A key consideration in developing a CDM is how much, if any, preprocessing of the information should be done upfront, and how much should be handled when conducting the study. Preserving the fidelity and granularity of the source information, as briefly discussed already, allows researchers of specific studies to determine the most appropriate study parameters. The extra time to execute the study, due to the additional deliberation, is generally worthwhile. Preprocessing the information upstream via mapping or creating specific constructs or concepts helps expedite the implementation of specific studies downstream, but may restrict researchers' ability to develop study parameters tailored to the studies. The approach taken by a given DDN depends on its missions, objectives,

and preferences. For example, the Sentinel system does minimal preprocessing of the data upfront, which allows the system to tailor its analysis to the FDA's regulatory questions. The OMOP CDM involves a significant amount of preprocessing, which allows researchers to streamline the conduct of their analysis using predefined variables and parameters. The pros and cons of these approaches, as well as their comparative performance, have been covered in the literature [75–78].

Methodological Advances

The defining feature of DDNs is that the data are stored locally under the direct control of participating data partners, and ideally only minimal necessary information is shared in each analysis. Traditionally, it had been necessary or preferred to share de-identified individual-level datasets for centralized analysis. With this approach, the participating sites send the analysis center an individual-level analytic dataset with distinct covariate information necessary for the analysis, yielding what is essentially a single centralized dataset after pooling. The confounders can be incorporated into the analysis through matching, stratification, restriction, regression, or weighting, and the data can be considered all together or stratified by contributing site [79,80]. Confounder summary scores (discussed shortly) can be estimated after centralizing the data. Although this approach offers the most analytic flexibility, it requires the most granular information among all the analytic options.

Recent methodological advances have expanded the data-sharing and analytic options, some of which require less granular information to perform the same type of analysis afforded by pooled individual-level data [81–85]. As a result, these newer methods may be preferred because they are more privacy protecting. Another feature of these new methods is that most or all the analyses will need to be specified *a priori*, or additional data requests may be required to obtain the additional information needed for *ad hoc* analyses. Although these are often seen as the limitations of these newer methods, they can also be considered strengths, because there is better transparency in the analysis. They help ensure clear delineation between prespecified and *ad hoc* analyses, and minimize opportunities for conducting unspecified analyses and selective reporting of results.

Individual-Level Confounder Summary Score-Based Methods

Confounder summary scores, such as propensity scores [86,87] and disease risk scores [88,89], are widely used in pharmacoepidemiologic research. If estimated correctly, these summary scores contain sufficient information to account for the confounding effects of the covariates used to estimate them. These data dimension reduction techniques have some appealing features useful for DDNs. Specifically, they obscure the information from a large number of covariates into scalar measures that are much less identifiable. Instead of requesting an individual-level dataset with information on individual covariates, one can replace these covariates with the summary scores [82,90,91]. In its simplest form, the dataset will only include variables indicating the treatment, outcome, follow-up (for time-to-event analysis), and confounder summary score. Other variables needed for the analysis, such as age or age categories if one wishes to perform age-stratified analysis, can also be requested. Conventional approaches to handling confounders, including matching, stratification, restriction, regression, and weighting, can then be done with the pooled, less granular, individual-level datasets.

This approach can perform essentially all the prespecified analyses afforded by the approach that shares individual confounder information, but it may not be able to accommodate all *ad hoc* analyses. For example, if sex is included in the estimation of the confounder summary score but is not requested separately, one will not be

able to perform a secondary, sex-stratified analysis without going back to the sites to request additional sex information.

The confounder summary scores should ideally be estimated and adjusted within each database, not just for practical reasons but also to ensure validity. For example, propensity scores are a function of the prevalence of the exposure in the population in which the scores are estimated. The prevalence of the exposure may vary across databases due to differences in formulary, regional prescribing pattern, and patient characteristics. Having site-specific propensity score models also allows the effect of a given covariate (e.g., age) on the probability of receiving the treatment of interest to vary by site. Researchers should account for site or database in the analysis by either including it as a stratification variable or performing within-site matching or stratification. As already discussed, data-adaptive approaches like the high-dimensional propensity score method [23] readily allow the analysis to be more tailored to the data availability at each participating site.

In contrast, the influence of risk factors on the outcome is generally more stable across databases, even if the outcome incidence varies by site. For example, the relation between age and heart failure, conditional on all other risk factors, should be similar across data sources. Therefore, it may be possible to combine disease risk scores, another commonly used confounder summary score that models and summarizes the associations between potential confounders and outcome risk, across sites. Additional research is needed to evaluate this issue.

Confounder Summary Score-Based Methods

It is possible to combine confounder summary score-based methods with other analytic techniques to further reduce the granularity of information shared. One can perform matching and stratification at the sites, and then only request the aggregate-level matched or stratified data to return to the analysis center [82,91]. In a matched analysis, if each site matches in the same fixed ratio, the only information needed for the analysis will be the total exposed and unexposed persons or person-times, and the number of exposed and unexposed outcomes. In a stratified analysis, participating sites send to the analysis center the total exposed and unexposed persons or person-times, and the number of exposed and unexposed outcomes within each stratum. Alternatively, one can structure the datasets into a risk set format at the sites and request risk set-based summary-level information for centralized analysis [82–84,92]. Results from the risk set-based approach have been shown to be statistically equivalent to results from the pooled individual-level stratified Cox regression model [83,92]. As with other methods, subgroup and sensitivity analyses will need to be prespecified so that appropriate summary-level information can be generated at the sites and shared for centralized analysis. As before, care should be taken with subgroup analyses to avoid small cells and potential identification risks. These methods protect against patient identification to an extent, but are not foolproof. If, for example, a rare disease or exposure is of interest, and prespecified analyses involve substantial stratification, some cells of the summary tables may be small enough to violate data partners' privacy regulations.

Meta-analysis of Database-Specific Results

An alternative to pooling individual-level data in a central repository is the commonly used approach of pooling site-specific effect estimates using meta-analytic techniques. In this approach, each site performs its own analysis, and the effect estimates and their variances (or other information needed to calculate database-specific weights) are provided to a central location and combined via meta-analysis [83,84,93–95]. The site-specific estimates can be obtained from matching, stratification, restriction, outcome modeling, or weighting, with or without confounder summary scores. This has

been shown to produce similar pooled effect estimates when compared with individual-level data analysis [90,96,97]. Although all data-sharing methods, including those discussed in this section, can in principle inspect treatment effect heterogeneity across databases, meta-analysis does that in the most obvious way, because database-specific effect estimates are shared and used in the pooled analysis. Each subgroup or sensitivity analysis requires all sites to perform each analysis internally, and then transfer the effect estimates to the lead team. Smaller sites may not be able to perform certain analyses, although sometimes using confounder summary scores to obtain site-specific effect estimate may help.

Distributed regression

The basic idea of distributed regression is for each data source to process its own individual-level data and share with the analysis center only summary statistics (e.g., sums of squares and cross-products matrix), such that the analysis center can either calculate the effect estimates or, if an iterative process is needed, update the parameter estimates and send them back to each data source to further update the summary statistics [98–101]. The iterative process continues until either a specified convergence criterion is met and the final parameter estimates are calculated, or the maximum number of iterations is reached. In other words, distributed regression conducts the same numeric algorithm with only centrally combined summary statistics as standard regression with pooled individual-level data. Although distributed regression is appealing in theory, it is relatively cumbersome to implement in practice, particularly for regression models that require multiple iterations. There are ongoing efforts to improve the practicality of distributed regression in existing DDNs [102–105].

Encryption

When applied in pharmacoepidemiology, encryption or hashing techniques are generally used to obscure potentially identifiable information while allowing valid database linkages [106,107]. In principle, it is possible to use these techniques to process the de-identified analytic dataset at the site before the encrypted data, along with the decryption method, are shared centrally for analysis. There have been some efforts in combining homomorphic encryption techniques with distributed regression [103,108], but using encryption to obscure potentially identifiable individual-level data is still quite theoretical and has not been widely implemented in practice.

The Future

More Sustainable and Efficient

DDNs, regardless of their actual configuration, typically require significant upfront investment. An architecture with a CDM is costly and time consuming to set up and maintain; however, efficiency benefits may be realized if the infrastructure is used for multiple studies. Some DDNs may achieve "economies of scale" and be able to conduct additional studies more efficiently and at a marginal cost compared to doing these studies as a series of "one-offs" [3,109]. DDNs without a CDM require infrastructure and initial investment to develop replicable and systematic processes. With more sustained funding support from regulatory agencies and other stakeholders, existing and future DDNs will have the much-needed stable foundation to grow, expand, and mature. This will be particularly important for some DDNs that are currently not supported by a single regulatory authority, such as AsPEN, and the various EU networks like PROTECT and EU-ADR. Even for funded DDNs, it can be argued that the infrastructure should be made available to other stakeholders with proper governance in place. The Sentinel system is an example of how this can be possible. Although the surveillance system was originally created by the FDA to meet its regulatory mandate, the agency envisioned the infrastructure eventually becoming a national resource for evidence generation [19]. Through the Innovation in Medical

Evidence Development and Surveillance initiative, non-FDA funders, including life science companies, can now access the same data sources and analytic tools used in Sentinel to conduct their own studies [110].

Broader Scope

With few exceptions, most DDNs were initially created for specific purposes (e.g., medical product safety surveillance). However, the scientific and technical infrastructure of many DDNs has the potential to address a wider range of topics, including comparative effectiveness research, patient-centered outcomes research, public health surveillance, and quality improvement. In addition to facilitating observational studies that analyze secondary data sources, some DDNs also support intervention studies. For example, PCORnet is conducting a pragmatic trial within participating health systems to examine aspirin dosing and secondary prevention of atherosclerotic cardiovascular disease [47]. Another pragmatic trial is underway in Sentinel to investigate the effect of direct mailings to patients and providers on initiation of anticoagulation therapy among eligible, treatment-naive patients [111]. These trials leverage the existing electronic healthcare databases of participating delivery systems or health plans to identify eligible patients and collect follow-up data, which allow the trials to be conducted more efficiently in real-world clinical settings compared to conventional randomized controlled trials.

More Diverse and Complementary Data Sources

Most DDNs are "horizontally partitioned," meaning that each database in the network contains information from different patients. Information is increasingly and routinely collected in various databases, for instance administrative claims databases, EHRs, disease or product registries, and data warehouses that contain information collected from wearables, mobile devices, or social media. In the future, we will likely see more DDNs that include disparate databases that include various data elements from the same individuals. These methods will require continued methodologic development to account for variation in data quality and completeness across data sources. Missing data and potential selection bias that arises from restricting the analysis to only patients appearing in multiple databases will require special attention.

More Robust and Secure Analysis

Continued methodologic advancement, both in developing cutting-edge analytic methods and in refining existing methods, will offer more analytic options that allow researchers to perform sophisticated statistical analysis while offering sufficient protection for patient confidentiality and data security. Existing methods already allow researchers to perform multivariable-adjusted outcome regression analysis and confounder summary-based analysis without sharing individual-level data for one-time exposures and one-time outcomes [82–84]. As DDNs mature, additional methodologic developments are likely to become available, including analysis of time-varying exposures, time-varying or repeated outcomes, missing data, and multilevel data.

Greater Transparency and Reproducibility

The DDN structure often requires researchers to prespecify the study design and analysis plan in advance, because they may not have direct access to all the data. For DDNs that employ a CDM, the analytic code will be pretested to ensure successful execution across the participating data partners. For DDNs that use a common protocol approach, the protocol will also need to be developed in advance. These data models, protocols, analytic code, and results should be made publicly available whenever

possible to improve transparency and encourage reproducibility [112]. Several DDNs, such as Sentinel and CNODES, are already adopting this policy, which has allowed other researchers to replicate the analyses [113,114]. OHDSI also makes its analytic tools publicly available.

Better Interoperability and Coordination across Networks

A possible future is one that has a national or international infrastructure that supports multiple DDNs. A healthcare delivery system or health plan can participate in multiple networks, each created for different purposes (e.g., medical product safety surveillance, comparative effectiveness research, pragmatic trials, public health surveillance). Each network will have its own governance and coordination. The networks can share the infrastructure, analytic tools, lessons learned, and software development and improvement. Regulators and decision-makers may also choose to collaborate on questions or work with multiple networks on specific queries.

References

1 AsPEN collaborators; Andersen M, Bergman U, Choi NK, *et al*. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf* 2013; **22**: 700–4.

2 Suissa S, Henry D, Caetano P, *et al*.; Canadian Network for Observational Drug Effects. CNODES: the Canadian Network for Observational Drug Effect Studies. *Open Med* 2012; **6**: e134–40.

3 Steiner JF, Paolino AR, Thompson EE, Larson EB. Sustaining research networks: the twenty-year experience of the HMO research network. *EGEMS (Wash DC)* 2014; **2**: 1067.

4 Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014; **21**: 578–82.

5 Reynolds RF, Kurz X, de Groot MC, *et al*. The IMI PROTECT project: purpose, organizational structure, and procedures. *Pharmacoepidemiol Drug Saf* 2016; **25** (Suppl 1): 5–10.

6 Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative: a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther* 2016; **99**: 265–8.

7 McNeil MM, Gee J, Weintraub ES, *et al*. The Vaccine Safety Datalink: successes and challenges monitoring vaccine safety. *Vaccine* 2014; **32**: 5390–8.

8 Lai EC, Man KK, Chaiyakunapruk N, *et al*. Brief report: databases in the Asia-Pacific region: the potential for a distributed network approach. *Epidemiology* 2015; **26**: 815–20.

9 Pratt N, Andersen M, Bergman U, *et al*. Multi-country rapid adverse drug event assessment: the Asian Pharmacoepidemiology Network (AsPEN) antipsychotic and acute hyperglycaemia study. *Pharmacoepidemiol Drug Saf* 2013; **22**: 915–24.

10 Pratt N, Chan EW, Choi NK, *et al*. Prescription sequence symmetry analysis: assessing risk, temporality, and consistency for adverse drug reactions across datasets in five countries. *Pharmacoepidemiol Drug Saf* 2015; **24**: 858–64.

11 Roughead EE, Chan EW, Choi NK, *et al*. Variation in association between thiazolidinediones and heart failure across ethnic groups: retrospective analysis of large healthcare claims databases in six countries. *Drug Saf* 2015; **38**: 823–31.

12 Roughead EE, Chan EW, Choi NK, *et al*. Proton pump inhibitors and risk of *Clostridium difficile* infection: a multi-country study using sequence symmetry analysis. *Expert Opin Drug Saf* 2016; **15**: 1589–95.

13 Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010; **153**: 600–6.

14 Chen RT, Glasser JW, Rhodes PH, *et al*.; Vaccine Safety Datalink Team. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. *Pediatrics* 1997; **99**: 765–73.

15 Chubak J, Ziebell R, Greenlee RT, *et al*. The Cancer Research Network: a platform for epidemiologic and health services research on cancer prevention, care, and outcomes in large, stable populations. *Cancer Causes Control* 2016; **27**: 1315–23.

16 Lu CY, Zhang F, Lakoma MD, *et al*. Changes in antidepressant use by young people and suicidal behavior after FDA warnings and media coverage: quasi-experimental study. *BMJ* 2014; **348**: g3596.

17 Go AS, Magid DJ, Wells B, *et al*. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. *Circ Cardiovasc Qual Outcomes* 2008; **1**: 138–47.

18 Ross TR, Ng D, Brown JS, *et al*. The HMO research network virtual data warehouse: a public data model to support collaboration. *EGEMS (Wash DC)* 2014; **2**: 1049.

19 Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel system: a national resource for evidence development. *N Engl J Med* 2011; **364**: 498–9.

20 Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new Sentinel network: improving the evidence of medical-product safety. *N Engl J Med* 2009; **361**: 645–7.

21 Platt R, Carnahan RM, Brown JS, *et al*. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012; **21**(Suppl 1): 1–8.

22 Gagne JJ, Han X, Hennessy S, *et al*. Successful comparison of US Food and Drug Administration Sentinel analysis tools to traditional approaches in quantifying a known drug-adverse event association. *Clin Pharmacol Ther* 2016; **100**: 558–64.

23 Zhou M, Wang SV, Leonard CE, *et al*. Sentinel modular program for propensity score-matched cohort analyses: application to glyburide, glipizide, and serious hypoglycemia. *Epidemiology* 2017; **28**: 838–46.

24 Connolly JG, Wang SV, Fuller CC, *et al*. Development and application of two semi-automated tools for targeted medical product surveillance in a distributed data network. *Curr Epidemiol Rep* 2017; **4**: 298–306.

25 Gagne JJ, Houstoun M, Reichman ME, Hampp C, Marshall JH, Toh S. Safety assessment of niacin in the US Food and Drug Administration's Mini-Sentinel system. *Pharmacoepidemiol Drug Saf* 2018; **27**: 30–7.

26 Davis RL, Kolczak M, Lewis E, *et al*. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology* 2005; **16**: 336–41.

27 Yih WK, Kulldorff M, Fireman BH, *et al*. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. *Pediatrics* 2011; **127**(Suppl 1): S54–S64.

28 Fahey KR. The pioneering role of the Vaccine Safety Datalink Project (VSD) to advance collaborative research and distributed data networks. *EGEMS (Wash DC)* 2015; **3**: 1195.

29 Lieu TA, Kulldorff M, Davis RL, *et al*. Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care* 2007; **45**: S89–S95.

30 Greene SK, Kulldorff M, Yin R, *et al*. Near real-time vaccine safety surveillance with

partially accrued data. *Pharmacoepidemiol Drug Saf* 2011; **20**: 583–90.
31 Madigan D, Ryan PB, Schuemie M, *et al*. Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol* 2013; **178**: 645–51.
32 Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med* 2012; **31**: 4401–15.
33 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.
34 Coloma PM, Schuemie MJ, Trifiro G, *et al*. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR project. *Pharmacoepidemiol Drug Saf* 2011; **20**: 1–11.
35 Arfe A, Scotti L, Varas-Lorenzo C; Safety of Non-steroidal Anti-inflammatory Drugs Project. Non-steroidal anti-inflammatory drugs and risk of heart failure in four European countries: nested case-control study. *BMJ* 2016; **354**: i4857.
36 CORDIS. Arrhythmogenic Potential of Drugs (ARITMO): project information. 2017. https://cordis.europa.eu/project/rcn/94061/factsheet/en (accessed May 2019).
37 CORDIS. Safety Evaluation of Adverse Reactions in Diabetes (SAFEGUARD): project information. https://cordis.europa.eu/project/rcn/100121/reporting/en (accessed May 2019).
38 CORDIS. Global Research in Paediatrics (GRIP): project information. https://cordis.europa.eu/project/rcn/97619/factsheet/en (accessed May 2019).
39 Accelerated Development of Vaccine Benefit-risk Collaboration in Europe (ADVANCE). About ADVANCE. http://www.advance-vaccines.eu (accessed April 2018).
40 Shin JY, Roughead EE, Park BJ, Pratt NL. Cardiovascular safety of methylphenidate among children and young people with attention-deficit/hyperactivity disorder (ADHD): nationwide self controlled case series study. *BMJ* 2016; **353**: i2550.
41 Dormuth CR, Hemmelgarn BR, Paterson JM, *et al*.; Canadian Network for Observational Drug Effects. Use of high potency statins and rates of admission for acute kidney injury: multicenter, retrospective observational analysis of administrative databases. *BMJ* 2013; **346**: f880.
42 Filion KB, Azoulay L, Platt RW, *et al*. A multicenter observational study of incretin-based drugs and heart failure. *N Engl J Med* 2016; **374**: 1145–54.
43 Henry D, Dormuth C, Winquist B, *et al*. Occurrence of pregnancy and pregnancy outcomes during isotretinoin therapy. *CMAJ* 2016; **188**: 723–30.
44 Graham DJ, Staffa JA, Shatin D, *et al*. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA* 2004; **292**: 2585–90.
45 Davis RL, Rubanowice D, McPhillips H, *et al*. Risks of congenital malformations and perinatal events among infants exposed to antidepressant medications during pregnancy. *Pharmacoepidemiol Drug Saf* 2007; **16**: 1086–94.
46 Habel LA, Cooper WO, Sox CM, *et al*. ADHD medications and risk of serious cardiovascular events in young and middle-aged adults. *JAMA* 2011; **306**: 2673–83.
47 Hernandez AF, Fleurence RL, Rothman RL. The ADAPTABLE Trial and PCORnet: shining light on a new research paradigm. *Ann Intern Med* 2015; **163**: 635–6.
48 Block JP, Bailey LC, Gillman MW, *et al*.; Antibiotics and Childhood Growth Study. The PCORnet Antibiotics and Childhood Growth Study: process for cohort creation and cohort description. *Acad Pediatr* 2018; **18**(5): 569–76.

49 Toh S, Rasmussen-Torvik LJ, Harmata EE, *et al.*; Collaborative PCBS. The National Patient-Centered Clinical Research Network (PCORnet) bariatric study cohort: rationale, methods, and baseline characteristics. *JMIR Res Protoc* 2017; **6**: e222.

50 Brauer R, Douglas I, Garcia Rodriguez LA, *et al*. Risk of acute liver injury associated with use of antibiotics: comparative cohort and nested case-control studies using two primary care databases in Europe. *Pharmacoepidemiol Drug Saf* 2016; **25**(Suppl 1): 29–38.

51 Souverein PC, Abbing-Karahagopian V, Martin E, *et al*. Understanding inconsistency in the results from observational pharmacoepidemiological studies: the case of antidepressant use and risk of hip/femur fractures. *Pharmacoepidemiol Drug Saf* 2016; **25**(Suppl 1): 88–102.

52 Schuerch M, Gasse C, Robinson NJ, *et al*. Impact of varying outcomes and definitions of suicidality on the associations of antiepileptic drugs and suicidality: comparisons from UK Clinical Practice Research Datalink (CPRD) and Danish National Registries (DNR). *Pharmacoepidemiol Drug Saf* 2016; **25**(Suppl 1):142–55.

53 Toh S, Reichman ME, Graham DJ, *et al.*; Mini-Sentinel Saxagliptin AMISWG. Prospective postmarketing surveillance of acute myocardial infarction in new users of saxagliptin: a population-based study. *Diabetes Care* 2018; **41**: 39–48.

54 Go AS, Singer DE, Toh S, *et al*. Outcomes of dabigatran and warfarin for atrial fibrillation in contemporary practice: a retrospective cohort study. *Ann Intern Med* 2017; **167**: 845–54.

55 Yih WK, Lieu TA, Kulldorff M, *et al*. Intussusception risk after rotavirus vaccination in U.S. infants. *N Engl J Med* 2014; **370**: 503–12.

56 Thompson WW, Price C, Goodson B, *et al*. Early thimerosal exposure and neuropsychological outcomes at 7 to 10 years. *N Engl J Med* 2007; **357**: 1281–92.

57 Lee GM, Greene SK, Weintraub ES, *et al.*; Vaccine Safety Datalink Project. H1N1 and seasonal influenza vaccine safety in the Vaccine Safety Datalink project. *Am J Prev Med* 2011; **41**: 121–8.

58 Gee J, Sukumaran L, Weintraub E; Vaccine Safety Datalink Project. Risk of Guillain-Barré syndrome following quadrivalent human papillomavirus vaccine in the Vaccine Safety Datalink. *Vaccine* 2017; **35**: 5756–8.

59 Maro JC, Brown JS, Dal Pan GJ, Li L. Orphan therapies: making best use of postmarket data. *J Gen Intern Med* 2014; **29**(Suppl 3): S745–51.

60 Kesselheim AS, Gagne JJ. Strategies for postmarketing surveillance of drugs for rare diseases. *Clin Pharmacol Ther* 2014; **95**: 265–8.

61 Kothari C, Wack M, Hassen-Khodja C, *et al*. Phelan-McDermid syndrome data network: integrating patient reported outcomes with clinical notes and curated genetic reports. *Am J Med Genet B Neuropsychiatr Genet* 2018; **177**(7): 613–24.

62 Sejvar JJ, Baughman AL, Wise M, Morgan OW. Population incidence of Guillain-Barré syndrome: a systematic review and meta-analysis. *Neuroepidemiology* 2011; **36**: 123–33.

63 Toh S, Hampp C, Reichman ME, *et al*. Risk for hospitalized heart failure among new users of saxagliptin, sitagliptin, and other antihyperglycemic drugs: a retrospective cohort study. *Ann Intern Med* 2016; **164**: 705–14.

64 European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Characterising the risk of major bleeding in patients with non-valvular atrial fibrillation: non-interventional study of patients taking direct oral anticoagulants in the EU. http://www.encepp.eu/encepp/viewResource.htm?id=21042 (accessed April 2018).

65 DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. *Stat Med* 1994; **13**: 1341–52; discussion 53–6.

66 Lan KKG, DeMets DL. Discrete sequential boundries for clinical trials. *Biometrika* 1983; **70**: 639–63.

67 Lan KK, DeMets DL. Changing frequency of interim analysis in sequential monitoring. *Biometrics* 1989; **45**: 1017–20.

68 Chrischilles EA, Gagne JJ, Fireman B, *et al*. Prospective surveillance pilot of rivaroxaban safety within the US Food and Drug Administration Sentinel System. *Pharmacoepidemiol Drug Saf* 2018; **27**: 263–71.

69 Mazor KM, Richards A, Gallagher M, *et al*. Stakeholders' views on data sharing in multicenter studies. *J Comp Eff Res* 2017; **6**(6): 537–47.

70 Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010; **48**: S45–S51.

71 Diamond CC, Mostashari F, Shirky C. Collecting and sharing data for population health: a new paradigm. *Health Aff (Millwood)* 2009; **28**: 454–66.

72 Curtis LH, Weiner MG, Boudreau DM, *et al*. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf* 2012; **21**(Suppl 1): 23–31.

73 PCORnet. PCORnet Common Data Model (CDM). 2018. http://www.pcornet.org/pcornet-common-data-model (accessed April 2018).

74 Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012; **19**: 54–60.

75 Bell C, Chakravarty A, Gruber S, *et al*. Characteristics of study design and elements that may contribute to the success of electronic safety monitoring systems. *Pharmacoepidemiol Drug Saf* 2014; **23**: 1223–5.

76 Gruber S, Chakravarty A, Heckbert SR, *et al*. Design and analysis choices for safety surveillance evaluations need to be tuned to the specifics of the hypothesized drug-outcome association. *Pharmacoepidemiol Drug Saf* 2016; **25**: 973–81.

77 Xu Y, Zhou X, Suehs BT, *et al*. A comparative assessment of Observational Medical Outcomes Partnership and Mini-Sentinel common data models and analytics: implications for active drug safety surveillance. *Drug Saf* 2015; **38**: 749–65.

78 Gagne JJ. Common models, different approaches. *Drug Saf* 2015; **38**: 683–6.

79 Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*, 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.

80 Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006; **60**: 578–86.

81 Rassen JA, Moran J, Toh D, *et al*. Evaluating strategies for data sharing and analyses in distributed data settings. 2013. https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_Methods_Evaluating-Strategies-for-Data-Sharing-and-Analyses_0.pdf (accessed April 2018).

82 Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. *Med Care* 2013; **51**: S4–S10.

83 Toh S, Shetterly S, Powers JD, Arterburn D. Privacy-preserving analytic methods for multisite comparative effectiveness and patient-centered outcomes research. *Med Care* 2014; **52**: 664–8.

84 Toh S, Reichman ME, Houstoun M, *et al*. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. *Pharmacoepidemiol Drug Saf* 2013; **22**: 1171–7.

85 Toh S, Platt R. Is size the next big thing in epidemiology? *Epidemiology* 2013; **24**: 349–51.

86 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.

87 Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79**: 516–24.

88 Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008; **95**: 481–8.

89 Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res* 2009; **18**: 67–80.

90 Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010; **19**: 848–57.

91 Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf* 2010; **19**: 858–68.

92 Fireman B, Lee J, Lewis N, Bembom O, van der Laan M, Baxter R. Influenza vaccination and mortality: differentiating vaccine effects from bias. *Am J Epidemiol* 2009; **170**: 650–6.

93 Platt RW, Dormuth CR, Chateau D, Filion K. Observational studies of drug safety in multi-database studies: methodological challenges and opportunities. *EGEMS (Wash DC)* 2016; **4**: 1221.

94 Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*, Version 501, updated September 2008. London: Cochrane Collaboration, 2008, Chapter 9.

95 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**: 177–88.

96 Toh S, Reichman ME, Houstoun M, *et al.* Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Arch Intern Med* 2012; **172**: 1582–9.

97 Li X, Fireman BH, Curtis JR, *et al.* Validity of privacy-protecting analytical methods that use only aggregate-level information to conduct multivariable-adjusted analysis in distributed data networks. *Am J Epidemiol* 2019; **188**(4): 709–23.

98 Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. *J Comput Graph Stat* 2005; **14**: 263–79.

99 Karr AF, Fulp WJ, Vera F, Young SS, Lin X, Reiter JP. Secure, privacy-preserving analysis of distributed databases. *Technometrics* 2007; **49**: 335–45.

100 Fienberg SE, Fulp WJ, Slavković AB, Wrobel TA. "Secure" log-linear and logistic regression analysis of distributed databases. *Lect Notes Comput Sci* 2006; **2006**: 277–90.

101 Slavković AB, Nardi Y, Tibbits MM. Secure logistic regression of horizontally and vertically partitioned distributed databases. *Proceedings of Workshop on Privacy and Security Aspects of Data Mining*. Washington, DC: IEEE Computer Society Press, 2007, pp. 723–8.

102 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. *J Am Med Inform Assoc* 2015; **22**: 1187–95.

103 El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. *J Am Med Inform Assoc* 2012; **20**: 453–61.

104 Wolfson M, Wallace SE, Masca N, *et al.* DataSHIELD: resolving a conflict in contemporary bioscience – performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 2010; **39**: 1372–82.

105 Her QL, Malenfant JM, Malek S, *et al.* A query workflow design to perform automatable distributed regression analysis in large distributed data networks. *EGEMS (Wash DC)* 2018; **6**(1): 11.

106 Kho AN, Cashy JP, Jackson KL, *et al.* Design and implementation of a privacy preserving electronic health record linkage tool in

Chicago. *J Am Med Inform Assoc* 2015; **22**: 1072–80.

107 Bezin J, Duong M, Lassalle R, *et al*. The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf* 2017; **26**: 954–62.

108 Hall R, Fienberg SE, Nardi Y. Secure multiple linear regression based on homomorphic encryption. *J Off Stat* 2011; **27**: 669–91.

109 Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff (Millwood)* 2014; **33**: 1178–86.

110 Reagan-Udall Foundation for the Food and Drug Administration. Innovation in medical evidence development and surveillance. http://reaganudall.org/innovation-medical-evidence-development-and-surveillance (accessed April 2018).

111 Pokorney SD, Cocoros N, Al-Khalidi H, *et al*. Implementation of a randomized controlled trial to improve treatment with oral anticoagulants in patients with atrial fibrillation (IMPACT-AFib). 2017. https://www.sentinelinitiative.org/sites/default/files/IMPACT-AFib_Protocol_v3.pdf (accessed April 2018).

112 Wang SV, Schneeweiss S, Berger ML, *et al*., Joint I-ISTFoRWEiHCDM. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Pharmacoepidemiol Drug Saf* 2017; **26**: 1018–32.

113 Li X, Girman CJ, Ofner S, *et al*. Sensitivity analysis of methods for active surveillance of acute myocardial infarction using electronic databases. *Epidemiology* 2015; **26**: 130–2.

114 Simeone JC, Nordstrom BL, Appenteng K, Huse S, D'Silva M. Replication of Mini-Sentinel study assessing mirabegron and cardiovascular risk in non-Mini-Sentinel databases. *Drugs Real World Outcomes* 2018; **5**: 25–34.