Drug Safety (2022) 45:493–510
https://doi.org/10.1007/s40264-022-01158-3

REVIEW ARTICLE

Applying Machine Learning in Distributed Data Networks for Pharmacoepidemiologic and Pharmacovigilance Studies: Opportunities, Challenges, and Considerations

Jenna Wong[1] · Daniel Prieto-Alhambra[2,3] · Peter R. Rijnbeek[3] · Rishi J. Desai[4] · Jenna M. Reps[5] · Sengwee Toh[1]

Accepted: 13 February 2022

Abstract

Increasing availability of electronic health databases capturing real-world experiences with medical products has garnered much interest in their use for pharmacoepidemiologic and pharmacovigilance studies. The traditional practice of having numerous groups use single databases to accomplish similar tasks and address common questions about medical products can be made more efficient through well-coordinated multi-database studies, greatly facilitated through distributed data network (DDN) architectures. Access to larger amounts of electronic health data within DDNs has created a growing interest in using data-adaptive machine learning (ML) techniques that can automatically model complex associations in high-dimensional data with minimal human guidance. However, the siloed storage and diverse nature of the databases in DDNs create unique challenges for using ML. In this paper, we discuss opportunities, challenges, and considerations for applying ML in DDNs for pharmacoepidemiologic and pharmacovigilance studies. We first discuss major types of activities performed by DDNs and how ML may be used. Next, we discuss practical data-related factors influencing how DDNs work in practice. We then combine these discussions and jointly consider how opportunities for ML are affected by practical data-related factors for DDNs, leading to several challenges. We present different approaches for addressing these challenges and highlight efforts that real-world DDNs have taken or are currently taking to help mitigate them. Despite these challenges, the time is ripe for the emerging interest to use ML in DDNs, and the utility of these data-adaptive modeling techniques in pharmacoepidemiologic and pharmacovigilance studies will likely continue to increase in the coming years.

Key Points

Many opportunities exist for distributed data networks (DDNs) to use machine learning in pharmacoepidemiologic and pharmacovigilance studies; however, the practical data-related characteristics of DDNs also create unique challenges for applying machine learning.

In this review, we discuss various challenges that DDNs face when applying machine learning and present different approaches for addressing these challenges, including issues for consideration and examples of how real-world DDNs have addressed or are working to help mitigate these challenges.

The use of machine learning in DDNs is an emerging area of interest that holds much promise, and the utility of these data-adaptive modeling methods for enhancing pharmacoepidemiologic and pharmacovigilance studies will likely continue to increase in the coming years.

✉ Sengwee Toh
darren_toh@harvardpilgrim.org

[1] Division of Therapeutics Research and Infectious Disease Epidemiology (TIDE), Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute, 401 Park Drive, Suite 401, East Boston, MA 02215, USA

[2] Pharmaco- and Device Epidemiology, Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK

[3] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

[4] Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[5] Janssen Research & Development, LLC, Titusville, NJ, USA

1 Introduction

The digital revolution has led to a growing abundance and availability of electronic health data capturing real-world uses of and experiences with medical products [1]. Increasing access to amassing amounts of digital health data, including data from administrative claims databases, electronic health record (EHR) systems, and disease-specific or product-specific registries, has garnered much interest in their use for studies in pharmacoepidemiology and pharmacovigilance [2]. The traditional practice of having numerous groups around the world use single databases to accomplish similar tasks and address common questions about medical products can be made more efficient through well-coordinated multi-database studies. By leveraging more data, multi-database studies are capable of producing more precise and generalizable findings, and they are better suited to investigate rare exposures and outcomes, as well as heterogeneous treatment effects [3]. Multi-database studies also facilitate the identification of larger cohorts of exposed patients in a shorter period of time—a crucial capability when timely answers to important questions are needed but limited data exist, such as ensuing the approval of new medications or public health emergencies, such as the COVID-19 pandemic [4].

Although the pooling of individual-level data from different databases, especially from similar healthcare systems, into a centralized location is an instinctive approach for conducting multi-database studies, it is often impractical because of ethical, legal, logistical, and administrative barriers [3, 5]. These obstacles have led to the rise of distributed data networks (DDNs), where databases (comprised of any type of data) are not pooled centrally and data partners maintain full control over the physical storage and use of their data (Fig. 1). Given the sensitive nature of the information contained in electronic health data, a DDN approach represents a conceptually favorable, and in most circumstances, more feasible approach to conducting multi-database analyses [3, 5].

There is a strong interest to incorporate machine learning (ML) in pharmacoepidemiologic and pharmacovigilance activities within DDNs [6, 7]. This interest is well founded given that ML in healthcare has been successfully applied to diagnose pathologies from medical images [8], extract structured information from unstructured clinical notes [9, 10], construct dense representations of medical concepts [9, 11, 12], predict health outcomes [13], identify data-driven descriptors of illnesses and diseases (i.e., phenotypes) [9, 14], and enhance confounding control in pharmacoepidemiology [15, 16]. Given the substantial amount of electronic health data accessible through DDNs, these ML advancements could be valuable for enhancing the use of such bountiful data. However, the siloed structure of DDNs and the disparate and diverse nature of the databases they contain also create unique challenges for applying ML.

In this paper, we discuss opportunities, challenges, and considerations for applying ML in DDNs for pharmacoepidemiologic and pharmacovigilance studies. We first discuss major types of activities performed by these DDNs and the ways in which ML may be used to help accomplish these activities (Sect. 2). Next, we discuss practical data-related factors that influence how DDNs work in practice (Sect. 3). We then bring together these discussions and jointly consider how the opportunities for using ML in pharmacoepidemiologic and pharmacovigilance activities are affected by the practical data-related factors of DDNs, leading to a number of challenges. We discuss different approaches for addressing these challenges, including issues for consideration and examples of how real-world DDNs have addressed or are working to help mitigate these challenges (Sect. 4). Finally, we conclude the paper by summarizing our main observations and perspectives on the use of ML in DDNs for pharmacoepidemiologic and pharmacovigilance studies (Sect. 5).

Throughout our discussion, we use the term "machine learning" to refer to highly flexible and data-adaptive algorithms that can automatically learn complex associations in high-dimensional data with minimal human guidance (e.g.,

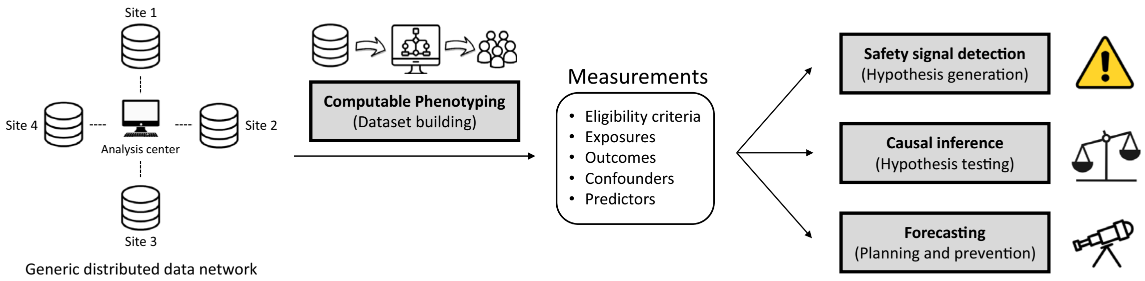

Fig. 1 Key domains of activities performed by distributed data networks conducting studies in pharmacoepidemiology and pharmacovigilance. On the left, a schematic of a generic distributed data network is shown, where data partners do not pool their databases and instead maintain full control over the use and sharing of their data with the analysis center. The gray rectangles represent key domains of activities performed by distributed data networks that conduct studies in pharmacoepidemiology and pharmacovigilance

random forests, support vector machines with Gaussian kernels, deep learning models) [17]. We focus our discussion on DDNs that conduct population-based pharmacoepidemiologic research, post-market medical product safety surveillance, or comparative effectiveness evaluations of medical products using databases of real-world data, in which the application of machine learning is particularly common and relevant. In the Box, we list select DDNs in pharmacoepidemiology and pharmacovigilance and refer interested readers elsewhere [3] for more details about these networks and their general characteristics.

1.1 Box. Examples of Distributed Data Networks in Pharmacoepidemiology and Pharmacovigilance

- Asian Pharmacoepidemiology Network (AsPEN) [18]
- Canadian Network for Observational Drug Effect Studies (CNODES) [19]
- European Health Data and Evidence Network (EHDEN) [20]
- Health Care Systems Research Network (HCSRN) [21]
- National Patient-Centered Clinical Research Network (PCORnet®) [22]
- Observational Health Data Sciences and Informatics (OHDSI) Collaborative [23]
- Sentinel System [24]
- Vaccine Safety Datalink [25]

2 Opportunities for Machine Learning

Many activities of DDNs that conduct studies in pharmacoepidemiology and pharmacovigilance fall into one of four key domains (Fig. 1). While some activities play a more intermediary role by supporting the creation of study cohorts and measurements, other activities generate hypotheses about medical product safety concerns, test hypotheses about medical products through addressing causal questions on effectiveness and safety, or identify high-risk individuals to inform planning and prevention efforts to improve patient outcomes and minimize harm from medical products. Below, we discuss each of these domains and highlight opportunities where ML may be used to help accomplish these activities.

2.1 Computable Phenotyping

Computable phenotyping (or simply "phenotyping") refers to the process of deriving computer-executable algorithms to identify individuals with specific health conditions, diseases, or clinical events based on measurable biological, behavioral, and clinical features [26]. Phenotyping activities are fundamental to the use of electronic health data in pharmacoepidemiology and pharmacovigilance [27], as they support essential tasks such as measuring study eligibility criteria, health outcomes of interest, confounders of treatment-outcome relations, and predictors of future health outcomes.

Phenotyping algorithms have traditionally consisted of expert-defined rules based on structured health data, such as medical codes or laboratory tests [26, 27]. Although interpretable and relatively simple to implement, rule-based phenotyping algorithms can only accommodate pre-existing knowledge or beliefs about a medical condition [26] and may be particularly challenging to develop when complex clinical criteria and tacit knowledge are required to make a diagnosis [28]. Furthermore, limiting phenotyping algorithms to using only structured data forgoes the opportunity to capitalize on the abundance of clinical information stored in unstructured data (e.g., clinical text), above and beyond that available in structured data, which are more challenging to extract but could be valuable for phenotyping activities [27].

In contrast, the use of ML for disease classification represents a more data-driven approach that can consider a multitude of clinical features to identify latent associations and potentially new phenotyping definitions. This process typically involves representing the information to be considered in the phenotyping algorithm as a feature vector, tagging a set of observations with labels (i.e., having or not having the phenotype of interest), and then allowing the data to train a supervised ML algorithm that maps the input features to the labels [9, 26]. Machine learning may also be used to facilitate the extraction of potentially relevant phenotypic information from clinical text as part of a natural language processing (NLP) tool, the output of which can then be used to identify phenotypes directly [29] or create NLP-derived features to be combined with other structured data in a downstream ML algorithm [27]. Phenotyping algorithms have been found to achieve better performance when developed using supervised ML compared to conventional decision rules [30] and when presented with features derived from both structured and unstructured data compared to structured data alone [31, 32]. In addition, ML has been used in other unique ways to enhance phenotyping activities, such as to estimate "probabilistic gold standard" phenotype probabilities on large groups of individuals to facilitate the efficient estimation of complete validation parameters (i.e., sensitivity and specificity, in addition to predictive values) for simple rule-based phenotyping algorithms commonly used for cohort development (e.g., one or more occurrences of a diagnosis code for the phenotype) [33].

2.2 Safety Signal Detection

Signal detection activities monitor the safety of medical products in real-world settings by detecting the emergence of new or unsuspected adverse events that may be associated with a product's use, where the identified signals are then further investigated for evidence of a potential causal association (discussed in Sect. 2.3). Compared to the other three domains of activities discussed in this section, safety signal detection activities are generally more specific to the field of pharmacoepidemiology and pharmacovigilance. Though signal detection activities have been traditionally performed using data from spontaneous reporting systems (SRSs), electronic health databases are being increasingly recognized as a valuable alternative data source because of their potential to address many limitations of SRS data (e.g., reporting bias, inability to estimate rates due to a lack of population denominators), among other reasons [34].

A variety of approaches for signal detection have been tested with electronic health databases, including approaches that transport methods originally designed for SRS data (e.g., disproportionality analyses), adopt traditional pharmacoepidemiologic designs (e.g., new-user cohort design, self-controlled case series), or utilize other methods such as the tree-based scan statistic, among many others [35, 36]. Within these approaches, several opportunities exist to use ML. For example, Bayesian Confidence Propagation Neural Networks [37] are used to estimate the Information Component (a well-known Bayesian disproportionality measure), and although traditionally used with SRS data, the approach has been tested with longitudinal (observational) data [35]. Machine learning may also be used to help reduce the effects of potential confounding in signal detection activities through the estimation of propensity scores (described later), which may be used with approaches such as the new-user cohort design or tree-based scan statistic [34, 38, 39]. In addition, other innovative approaches for using ML with longitudinal data in signal detection activities are being explored. For example, Reps et al. [40] proposed a supervised ML framework to predict the likelihood of a drug-event pair being an adverse drug reaction based on a vector of risk ratios calculated under different simple cohort study designs, where a ML classifier is trained on a sample of drug-event pairs known to be adverse drug reactions or not. There is also a growing interest to use ML methods, especially deep learning models, to extract mentions of drug-adverse event pairs from unstructured clinical text [41, 42]—an NLP task with the potential to improve the identification of adverse drug events that are typically under coded in structured EHR and claims data [43, 44].

2.3 Causal Inference

Among the most well-recognized pharmacoepidemiologic activities of DDNs, causal inference activities address important questions about the comparative safety and effectiveness of medical products to provide actionable evidence for informing clinical and public health decisions. A key component of causal inference is clearly defining the hypothesis or question of interest and statistically formulating a corresponding causal parameter that answers the question and can be validly estimated from available data [45].

When longitudinal data from electronic health databases are used to estimate causal parameters, the ability to control for imbalances in risk factors between treatment groups (i.e., confounding) is crucial and often achieved through estimating "nuisance functions" that are not of direct interest, but rather used as a means of estimating causal parameters [45–47]. Technically, nuisance functions estimate the probability of the treatment (called propensity scores) or outcome, conditional on a set of observed covariates that should follow sound epidemiologic principles (e.g., propensity score models should not include covariates associated with only the treatment and not the outcome [15, 48]). Depending on the causal effect estimation method, one or both nuisance functions may be used. For example, propensity score-based methods use the treatment nuisance function, G-computation methods use the outcome nuisance function, and doubly-robust methods such as targeted maximum likelihood estimation and augmented inverse probability weighting use both treatment and outcome nuisance functions [47]. Regardless of the method, at least one nuisance function must be properly specified to attain consistent estimates of causal parameters [47].

Parametric regression models are commonly used to estimate nuisance functions, but the strong functional assumptions they make are prone to misspecification, particularly when there are many covariates with potentially complex associations. Alternatively, data-adaptive supervised ML techniques offer a more flexible modeling approach for estimating nuisance functions that impose less restrictive assumptions on covariates, thus increasing the likelihood of attaining properly specified nuisance models for more valid inferences [15, 49], especially when used with cross-fitting estimation procedures [47]. Nuisance functions estimated using ensemble ML techniques [50], which can consider collections of different ML algorithms and different covariate sets (e.g., using various thresholds of top-ranked covariates from the high-dimensional propensity score algorithm [51]), have been shown to yield more consistent estimates of causal parameters [47, 52, 53]. Machine learning methods may also be useful for further automating the confounding adjustment process by efficiently extracting and prioritizing appropriate covariates from high-dimensional spaces in claims data, structured EHR data, and unstructured EHR data (e.g., word stems or N-grams) for inclusion in nuisance function models [54].

2.4 Forecasting

Forecasting activities predict the risk of future health events, outcomes, or behaviors (e.g., adverse drug events, treatment response, or nonadherence behaviors) to inform early intervention, planning, and prevention efforts, with the ultimate goal of improving patient outcomes, minimizing risks, and managing healthcare services [55]. Approaches to forecasting have been classified as judgmental or statistical [56]. Healthcare providers inherently use their expert judgment about the likelihood of future health events whenever they make decisions about treatments or recommend healthcare services for their patients. Though essential to the practice of medicine, judgment-based forecasting by clinicians is subjective and does not always translate into accurate predictions [57–59]. Moreover, it does not scale well to forecasting on large groups of individuals to provide healthcare delivery systems with the macro-level estimates needed for planning and management.

In contrast, statistical forecasting uses statistical models based on historical data to predict the occurrence of future health events (i.e., for prognostic modeling). Such prognostic models not only enable population-level forecasting, but they can also be used to further inform clinicians' judgments and enhance decision making at the patient level (e.g., as decision support tools). Similar to the use of supervised ML for diagnostic modeling in phenotyping activities, the use of supervised ML for prognostic modeling offers a data-adaptive approach for identifying potentially complex associations and interactions between a plethora of features to predict future health outcomes. Supervised ML has been used with electronic health data to create accurate risk prediction models for many health events for which prospective surveillance, prevention, early intervention, and advanced planning is invaluable, including opioid overdose [60, 61], cancer-related mortality [62], suicidality [63], and high healthcare costs [64].

3 Practical Data-Related Factors

In this section, we shift our attention to practical matters, focusing on three data-related factors that influence how DDNs perform their activities (Fig. 2). We discuss each of these factors and the "spectrum of possibilities" along which DDNs may fall, using examples of well-known DDNs to highlight their similarities and differences. Although other factors may also influence how DDNs carry out their activities (e.g., funding, experience working together, data infrastructure) [3], we focus on these data-related factors for their proximate influence on the application of ML in DDNs (discussed in Sect. 4).

3.1 Modality of Source Data

Electronic health data exist in formats that are either structured or unstructured [65]. Structured data have a well-defined format (e.g., tables), where information is often stored using standardized values and can be easily extracted and analyzed. Information commonly stored as structured data include patient demographics (e.g., age), medications, coded diagnoses and procedures, certain laboratory tests (e.g., international normalized ratio), and certain quantitative clinical measures (e.g., blood pressure). In contrast, unstructured data (e.g., clinical notes, discharge summaries, radiology and pathology reports, and medical images) contain valuable information beyond that captured in structured data, but their lack of any pre-defined structure makes them challenging for computers to process and analyze [65]. Depending on the source of electronic health data, one or both data modalities may exist. For example, administrative claims databases contain exclusively structured data, while

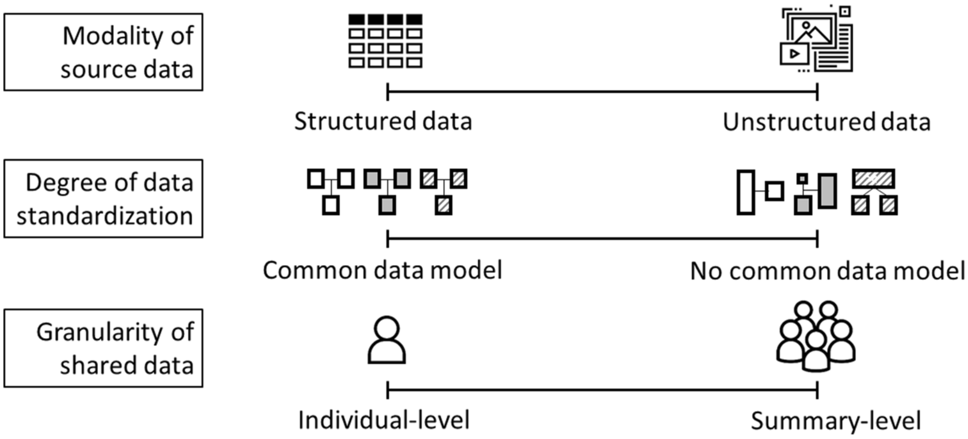

Fig. 2 Practical data-related factors of distributed data networks. These three data-related factors influence how distributed data networks operate in practice. For each factor, distributed data networks may theoretically fall anywhere along the spectrum between the two extremes

EHR databases typically contain a mixture of structured and unstructured data, with the majority of these data often being unstructured [66].

Currently, most DDNs in pharmacoepidemiology and pharmacovigilance primarily use structured data from their source systems to perform activities; thus, most DDNs are situated toward the leftward end of the spectrum in Fig. 2. For DDNs comprised solely of administrative claims databases, the use of only structured data occurs because unstructured data are not available. However, for many DDNs containing EHR databases amongst their data partners, such as the Sentinel System, National Patient-Centered Clinical Research Network (PCORnet®), Health Care Systems Research Network (HCSRN), European Health Data and Evidence Network (EHDEN), and Observational Health Data Sciences and Informatics (OHDSI) Collaborative, both structured and unstructured data often exist in these databases. However, the unstructured data in these databases have been traditionally underutilized, not because of a lack of interest, but rather owing to the significant complexities associated with handling and extracting information from these complex data types. Given the rapid advancements in NLP methods and growing literature on clinical NLP applications in recent years [67–69], there is increasing interest among many DDNs to move rightward along the spectrum and make greater use of the unstructured data that may exist in their source systems. For example, the Sentinel System, as part of its 5-year strategy, has committed to exploring emerging data science innovations, including NLP, to expand the use of EHR data in its activities [6], and the OHDSI collaborators have established an NLP working group that develops methods and software to promote the use of clinical text in activities within the OHDSI community [70].

3.2 Degree of Data Standardization

In their native environments, most electronic health databases vary greatly in their schemas, content, and coding terminologies [71]. Given the significant discrepancies that often exist between disparate databases in a DDN, a range of approaches may be used to deal with these differences when conducting a distributed analysis. At one end of the spectrum, DDNs can have all their data partners standardize the format of their source data to a common data model (CDM). A CDM specifies a standardized structure and set of tables to which data partners in a DDN convert their source data [3, 71]. Some CDMs, such as the Observational Medical Outcomes Partnership (OMOP) CDM, maintained by the OHDSI group and also used by EHDEN and AsPEN, additionally standardize coding terminologies to a common vocabulary [72], whereas other CDMs such as the Sentinel CDM only store the original coded values to be later mapped to a common vocabulary on a study-specific basis [73]. Although CDMs require significant up-front investment to both develop the infrastructure and implement the specifications, the invested time and effort pay increasing dividends as more studies are performed [3]. In particular, use of a CDM offers the ability to apply validated software and tools that have been developed for the CDM, thus promoting not only the rapid, reliable, and reproducible implementation of analyses across sites, but also reducing the amount of site-specific programming required and coding errors encountered [71]. The Sentinel System, PCORnet, and OHDSI are examples of DDNs that have developed libraries of customized tools for use with data formatted to their CDMs [3]. For example, members of the OHDSI community may utilize open-source R packages available in the Health Analytics Data-to-Evidence Suite (HADES) [74] to perform various study analyses on data converted to the OMOP CDM.

At the other end of the spectrum, DDNs may choose not to use a CDM. Rather than invest time and energy up-front to standardize their data partners' source systems, data management, quality, and harmonization issues are instead addressed at the study level. Often, this approach involves creating a meticulous data management plan along with a detailed statistical analysis plan representing components of a structured pre-specified protocol developed collaboratively between researchers and stakeholders involved in the study [73]. This approach was used by the historical Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium (IMI-PROTECT) project [75] and in the early years of the Canadian Network for Observational Drug Studies (CNODES) [19]. Intermediary approaches between these two extremes are also possible. For example, a CDM may be developed for only some data partners within a DDN, or only a small fraction of information within data partners' source systems may be converted to a CDM.

Currently, most DDNs conducting pharmacoepidemiologic and pharmacovigilance studies use a CDM to standardize primarily structured data from their data partners' source systems. Thus, most DDNs tend to lie more leftward along the spectrum in Fig. 2. Distributed data networks once located more rightward along the spectrum have also been inclined to shift leftward over time toward adopting a CDM, indicating the clear advantages of investing in a CDM when DDNs conduct increasing numbers of distributed analyses over time. For example, although CNODES initially used a phased common protocol approach without a CDM to conduct their studies, they recently launched an initiative to gradually transition their sites toward adopting a Sentinel-like CDM [7].

3.3 Granularity of Shared Data

Once all sites have prepared their analytic datasets containing the necessary study variables for all eligible individuals in their databases, there are different ways in which these individual-level datasets may be used to perform the activity at hand. At one end of the spectrum, each site in the network can share its individual-level dataset with the analysis center, which may also be a data-contributing site. Once the individual-level datasets have been pooled, the result is a single centralized dataset [3]. This process is distinct from a centralized data approach because only the final study datasets—not the underlying data sources from which they were curated—are shared, subject to each data partner having approved the request and reviewed its final dataset before sharing. Although this approach requires the most granular level of data sharing (and is thus the least privacy protecting), it offers the most analytic freedom, such that decisions about the methods used and how they are applied to the data can be based solely on the scientific needs of the study [76]. For example, if the analysis center felt it was best to fit a separate treatment nuisance model for each site but a global outcome nuisance model across all sites, then this plan could be easily implemented. To minimize privacy and confidentiality concerns, the individual-level datasets are typically de-identified (e.g., by removing personal identifiers), and if desired, extra care can be taken to reduce their dimensionality (e.g., by combining individual variables into summary measures that contain essentially the same information using fewer variables) [3]. In addition, the ability to share de-identified individual-level data generally requires proper governance, appropriate data use agreements, and established collaborative relationships between sites, leading to a shared sense of trust. Such elements are important pre-requisites that are necessary, though not always sufficient, to allow for sharing of individual-level data between entities [3].

At the other end of the spectrum, each site in the network can share only summary-level data with the analysis center. Different types of summary-level data may be shared depending on the type of activity being performed, the methodologic approach being used, and the degree of privacy protection desired. For example, when conducting causal inference activities while sharing only summary-level data, a stratified analysis can be implemented by having each site send aggregated counts of the total number of persons (or person-time) and outcomes in each treatment group per stratum to the analysis center, essentially representing a coarsened version of the individual-level dataset [76]. Another approach is to use distributed regression, which produces identical results to a centralized outcome regression analysis of individual-level data. Distributed regression can be implemented by having each site fit a local regression on its own data and share only intermediate model statistics (e.g., sums of squares and cross-products matrix) with the analysis center, which then calculates the global parameter estimates, and if necessary, sends them back to each site for additional processing to update the global parameters iteratively until a pre-specified convergence criterion is met [3]. A third possible approach, which offers the greatest amount of privacy protection, is to have each site conduct its own analysis and estimate a site-specific causal parameter estimate that can then be pooled by the analysis center via a meta-analysis [76]. Regardless of the approach, when sites share only summary-level data, the analysis plan is typically more constrained because it must consider not only the scientific needs of the study, but also the practical challenges of being unable to combine and process individual-level data from each site. In addition to sharing only individual-level or summary-level data, other data sharing combinations are also possible. For example, only some sites within a DDN may share individual-level data with the analysis center, or all sites may share individual-level data for some projects but only summary-level data for other projects.

Currently, DDNs that conduct pharmacoepidemiologic and pharmacovigilance studies are scattered across the spectrum of possibilities in Fig. 2. For example, data partners within PCORnet and HCSRN have been known to share de-identified individual-level data for some studies, while data partners within Sentinel, CNODES, OHDSI, and EHDEN generally share only summary-level data.

4 Challenges and Considerations

In this section, we combine our prior discussions and consider how the opportunities for using ML in pharmacoepidemiologic and pharmacovigilance studies (discussed in Sect. 2) are affected by practical data-related factors for DDNs (discussed in Sect. 3). To guide the discussion, we consider four select scenarios, each with different characteristics in terms of a DDN's location along the spectrum of data-related factors (Table 1). For each scenario, we discuss unique challenges that arise for DDNs during the ML process, as well as possible approaches for addressing these challenges and issues for consideration. We also highlight efforts that real-world DDNs have taken to help mitigate these challenges, focusing largely on initiatives within the Sentinel System and OHDSI community, and we describe examples of select studies where ML has been used in DDNs (Table 2).

4.1 Scenario 1: Base Case

We first consider a DDN located at the leftmost end of the spectrum for all three data-related factors in Fig. 2. In this scenario, all sites in the DDN utilize only structured data from their source systems, these structured data populate a CDM containing all information needed to create the inputs for an activity, and all sites share their final de-identified individual-level datasets with the analysis center. We refer to this scenario as the “base case”, which also serves as the reference for Scenarios 2–4.

From a technical perspective, the base case represents the most simple and straightforward setting for applying ML in DDNs. In the data preparation stage, the use of only structured data greatly facilitates the creation of a CDM, and the use of inputs derived entirely from standardized fields within the CDM facilitates the feature engineering process and curation of datasets for ML. In the model fitting stage, the ability to pool the site-specific datasets into a centralized dataset essentially allows the modeling process to proceed with the same flexibility as in a single database setting. Although DDNs under this scenario may still face technical challenges (e.g., missing data), the nature of these challenges will be comparable to those encountered in single database settings (albeit on a larger scale), and in theory, can be addressed using the same approaches as for single databases (e.g., imputation) [77].

From a scientific perspective, the base case still presents issues for consideration when applying ML because of heterogeneity that may exist between databases in a DDN. In other words, although it is technically possible to combine all the datasets and apply ML to the centralized dataset, *should* it be done? Heterogeneity between databases can exist for a variety of reasons, including differences in data encoding (e.g., data quality and coding practices) and content (e.g., available data elements and domains), as well as variations in patient characteristics and care. To help disentangle true heterogeneity in patient populations and care from data quality problems, it is important that DDNs conduct regular and robust data quality assessments, ideally according to a systematic and conceptually based framework [78–81]. For example, to minimize data quality issues and errors that may arise during the CDM creation process, the Sentinel System requires that all extracts from their data partners first pass an extensive data quality review process [73, 82]. The OHDSI collaborators have also developed the Data Quality Dashboard [83], representing an open-source tool that performs a series of systematic data quality checks on databases mapped to the OMOP CDM to report potential data quality issues before these databases are used in modeling activities [13]. To further identify and reduce potential heterogeneity in coding practices between data partners, the Sentinel System is currently exploring the use of novel code translation methods, which incorporate unsupervised learning and language translation methods, to generate data-driven code mappings that could be used as a scalable and automated approach to help address idiosyncrasies in the coding process and harmonize medical codes across databases [84, 85]. Such initiatives highlight the importance of preserving the original data values to the greatest extent possible when populating the CDM to maintain high fidelity and minimize information loss during the conversion process—a guiding principle of both the Sentinel and OMOP CDMs (the OMOP CDM both maps codes to a common vocabulary and retains the original source codes) [73, 86].

Ultimately, even if it is technically possible to create a centralized dataset across sites, the most appropriate approach for applying ML (e.g., analyzing the data all together or stratified by site) will depend on the purpose of the ML model and the extent of known or suspected heterogeneity across databases. In causal inference activities, as the prevalence of treatment and the impact of patient covariates on the probability of treatment often vary across databases, propensity scores are generally estimated using models that are stratified by database or flexible enough to allow for database-specific effects of covariates on the propensity score [76]. For example, in a multi-site study conducted within PCORnet [87], ML was used with the latter approach to calculate propensity scores for confounding control in assessing the comparative effectiveness and safety of different bariatric procedures (Table 2). In contrast, when estimating the outcome nuisance function in causal inference activities, it may be appropriate to fit a global model on all the data as it is reasonable to suspect that the impact of individual risk factors on the probability of the outcome may be more stable across databases [76]. Fortunately, when sites are able to share their individual-level datasets, the analysis

Table 1 Four select scenarios of distributed data networks

Scenario	Modality of source data	Degree of data standardization	Granularity of shared data
1 (base case)	Structured data only	Common data model for all inputs	Individual-level data for all sites
2 (less standardized data available)	Structured data only	**No common data model for some inputs**	Individual-level data for all sites
3 (more complex data modalities used)	**Structured and unstructured data**	Common data model for all inputs	Individual-level data for all sites
4 (less granular data shared)	Structured data only	Common data model for all inputs	**Summary-level data for all sites**

Scenario 1 represents the most simple and straightforward case for applying machine learning in distributed data networks. Scenarios 2–4 each deviate from the base case with respect to characteristics under one of the three practical data-related factors, as indicated in bold

Table 2 Select studies involving the use of machine learning in distributed data networks

Objective	Study setting	Use of machine learning	Main findings
To compare the effectiveness and safety of three bariatric procedures: RYGB, SG, and AGB [87]	41 US health systems in the National Patient-Centered Clinical Research Network ($n = 46{,}510$)	• LASSO used to simultaneously select features and estimate parameters for propensity score models • As de-identified individual-level datasets were shared with the analysis center, site-specific effects of covariates on propensity scores were allowed by including interactions between site and all covariates in the feature selection process for propensity score models • Propensity score deciles used for confounding adjustment in the association of bariatric procedure type with the study outcomes	• RYGB associated with greater weight loss than SG or AGB at 1-year, 3 years, and 5-years post-procedure, but RYGB had the highest 30-day rate of major adverse events • At 5 years, estimated percent total weight loss for RYGB patients was 6.7 (95% CI 5.8–7.7) percentage points greater than SG patients and 13.9 (95% CI 12.4–15.4) percentage points greater than AGB patients • 30-day rate of major adverse events was 5.0% for RYGB vs 2.6% for SG patients (OR 1.57, 95% CI 1.40–1.77) and 2.9% for AGB patients (OR 1.66, 95% CI 1.28–2.16)
To develop and validate a phenotyping algorithm for anaphylaxis [94]	KPWA ($n = 239$) and KPNW ($n = 277$)	• 5 machine learning algorithms (logistic regression, elastic net, BART, feed-forward neural network and boosted trees) used to predict the probability of being an anaphylaxis case • Ensemble learner containing a weighted combination of the machine learning algorithms also considered • Candidate features manually curated from structured EHR data and unstructured clinical text • 3 feature selection approaches explored • 3 feature sets explored to determine the added value of including features derived from unstructured text • All phenotyping algorithms developed and internally validated at KPWA; transported and externally validated at KPNW	• Adding features derived from unstructured clinical text improved the performance of phenotyping algorithms compared with using features from structured data alone • At KPWA, BART based on features from structured data and clinical text, selected using LASSO, achieved the best performance (cross-validated AUC: 0.71) • BART based on features from structured data and clinical text with no feature selection generalized best to KPNW (cross-validated AUC: 0.70 at KPWA, 0.67 at KPNW) • Manual curation of NLP-derived features was extremely labor intensive; future work will explore semi-automated approaches for curating features from clinical text
To develop and validate a prognostic model predicting risk of hemorrhagic transformation within 30 days of an initial acute ischemic stroke [106]	11 databases from three countries (USA, Germany, and Japan) within the OHDSI network ($n = 6{,}848{,}096$)	• LASSO used to simultaneously select features and estimate parameters to predict risk of hemorrhagic transformation within 30 days of initial acute ischemic stroke • Candidate features created from structured EHR data within 3 lookback periods prior to the index date • Prognostic model developed and internally validated in 1 database; externally validated in each of the remaining 10 databases	• Of 169,967 candidate predictors considered, 612 selected for the final model • In the development database ($n = 776{,}437$), internal validation of the final model had an AUC of 0.75 in the hold-out test set • Across the remaining 10 databases ($n = 6{,}071{,}659$), external validation of the final model achieved an AUC range from 0.60 to 0.78

Table 2 (continued)

Objective	Study setting	Use of machine learning	Main findings
To develop and validate a machine learning model to predict cause of death (within 60 days) from a patient's last medical check-up [105]	2 databases from South Korea with cause-of-death data ($n = 903{,}812$), 3 US databases with no cause-of-death data ($n = 2{,}397{,}494$)	• Two-level stacking ensemble used to predict cause of death within 60 days, which consisted of a meta-learner that used the outputs from a collection of base learners as inputs to make the final prediction • The base learners consisted of 2 machine learning algorithms (LASSO and gradient boosting machine) that predicted each of 9 outcomes (mortality status and 8 causes of death), for a total of 18 base learners • Candidate features created from claims data within 3 lookback periods prior to the index date • Stacking ensemble developed and internally validated in 1 South Korean database (claims); externally validated in the other South Korean database (EHR) • Stacking ensemble also used to impute cause of death in the 3 US databases	• In the development database ($n = 174{,}747$), internal validation of the stacking ensemble had an AUC of 0.95 in the hold-out test set; external validation of the stacking ensemble had an AUC of 0.89. • In 1 US database with mortality status, the AUC of the 2 base learners predicting mortality status were both 0.98, but the top 3 causes of death imputed by the stacking ensemble differed from the known top-ranked causes of mortality in the USA; these discrepancies suspected to be at least partly attributable to differences between the countries in the definition of heart disease death.
To develop and validate a prognostic model predicting 1-year risk of incident heart failure in patients with type 2 diabetes mellitus initiating a second pharmacotherapy for type 2 diabetes [108]	5 US databases ($n = 403{,}187$)	• LASSO used to simultaneously select features and estimate parameters to predict 1-year risk of incident heart failure • 2 feature sets evaluated (age and sex, all features) • Each database developed and internally validated 2 prognostic models (1 per feature set); remaining 4 databases externally validated the site-specific models	• Internal validation of site-specific models had an AUC range from 0.64 to 0.71 for baseline (age and sex) models and an AUC range from 0.73 to 0.81 for full models • Among full models, external validation of 3 site-specific models consistently achieved comparable performance across all other databases • Using a heatmap to visualize the internal and external performance of the site-specific models across all databases offers valuable insights
To predict clinical outcomes in patients with COVID-19 who present to the emergency department [113]	20 clinical institutes from various regions around the world ($n = 16{,}148$)	• Federated learning used to train a deep learning model to predict the EXAM (electronic medical record chest X-ray AI model) risk score, a continuous value from 0 to 1, with higher values denoting greater oxygen requirements • Model inputs included 20 features, 19 derived from the electronic medical record data and 1 chest x-ray image, where electronic medical record and image data were concatenated into a single high-dimensional feature vector • Minimal efforts directed at harmonizing data across sites • Global and local models trained; evaluated on held-out test data at each site	• For predicting 24-hour oxygen treatment, global model outperformed all local models: average AUC for the global vs locally trained models was 0.92 vs 0.80 (16% improvement), and AUC for global model also provided an average increase in generalizability of 38% compared with AUC for locally trained models • At the largest site, global model achieved sensitivity of 0.95 and specificity of 0.88 for predicting mechanical ventilation treatment or death at 24 hours

AGB adjustable gastric banding, *AUC* area under the receiver operating curve (possible values from 0 to 1, values closer to 1 indicate better performance), *BART* Bayesian additive regression trees, *CI* confidence interval, *EHR* electronic health record, *KPNW* Kaiser Permanente Northwest, *KPWA* Kaiser Permanente Washington, *LASSO* Least Absolute Shrinkage and Selection Operator, *NLP* natural language processing, *OHDSI* Observational Health Data Sciences and Informatics, *OR* odds ratio, *RYGB* Roux-en-Y gastric bypass, *SG* sleeve gastrectomy center has the freedom to experiment with various analytic options to better understand the potential influence of database heterogeneity on the study findings and select the most appropriate approach for the task at hand.

4.2 Scenario 2: Less Standardized Data Available

In this scenario, we consider a DDN with the same characteristics as the base case, but instead of the CDM containing all information needed to create the study inputs, some information of interest exists outside the CDM in the native (and thus unstandardized) structured data within data partners' source systems. When using data-adaptive ML methods, this situation may be encountered frequently because it is often of interest to consider a wide range of features that may not all be measurable from existing fields in the CDM.

When information of interest exists outside the CDM, the unstandardized data across sites creates challenges during the feature engineering process. One approach for addressing this challenge is to standardize the unstandardized information, which may be done by having sites create "data sidecars" for the study (i.e., tables containing additional data elements that can be linked to the primary tables in the CDM) [88] or by expanding the CDM itself to include new fields or tables. However, because these approaches are often costly and time consuming to implement, the anticipated benefits must be carefully weighed against the anticipated costs. For example, the HCSRN requires that any specification changes or table additions to its Virtual Data Warehouse CDM are first proposed by workgroups, then discussed by the Virtual Data Warehouse Implementation Group, and finally approved through a formal voting process [89]. Often, these efforts are only pursued if the information of interest can be easily obtained (low effort), will be frequently used (high yield), or is urgently required (high demand). For example, to support public health efforts during the COVID-19 pandemic, both PCORnet [90] and the Sentinel System [91] added new fields to their core CDMs to capture important information needed to better characterize infected patients. In addition, the Sentinel System, as part of its 5-year strategy [6], is looking to expand its CDM (currently mostly claims based) to include additional fields containing more granular clinical information from structured data fields in EHRs (e.g., vital signs and body mass index)—an initiative that will broaden the range of stakeholder questions the Sentinel System can consider in future projects.

Another approach for handling unstandardized information of interest outside the CDM is to allow sites to perform site-specific modeling with the additional (unstandardized) variables. Sites can also compare their results with and without the additional variables to see if their inclusion produces any meaningful changes. In many cases, this less resource-intensive approach may be more practical if some sites do not contain the additional data elements of interest and because when fitting data-adaptive ML models, it is often of interest to consider a large number of variables whose importance in the ML model, both individually and collectively, may be unknown during the feature engineering stage. Thus, it may not always be worth the extra resources to standardize additional variables unless one has strong reason to believe that their inclusion will produce an important impact on the model performance and results.

4.3 Scenario 3: More Complex Data Modalities Used

In this scenario, we again consider a DDN with the same characteristics as the base case, but instead of using only structured data from its data partners' source systems, it also uses unstructured data. For the purposes of this discussion, we only consider the use of unstructured text (e.g., clinical notes), but the ideas discussed here similarly apply to other types of unstructured clinical data (e.g., medical images). The richer clinical information stored in unstructured text provides the opportunity to enhance the performance and value of using data-adaptive ML methods in DDN activities. However, the more complex requirements for handling and processing unstructured text also create challenges for ML during the feature engineering process. In essence, one can view these challenges as an extension of the challenges in Scenario 2, where in this case, the information of interest outside the CDM is unstructured text.

Similar to the second approach described in Scenario 2, one practical approach is for sites to perform a site-specific analysis, such that all processing and information extraction that happens on the unstructured text occurs outside the CDM according to a pre-defined protocol. The Sentinel System followed this approach in a series of pilot projects using structured and unstructured EHR data to create a phenotyping algorithm for anaphylaxis [92, 93], where the use of features derived from unstructured clinical text was found to improve the performance of the phenotyping algorithm compared with using features from structured data alone—a finding that persisted even when the phenotyping algorithm (developed at one Sentinel site) was transported to a second Sentinel site [94] (Table 2). This pilot project is now guiding the development of a general framework for using ML and NLP techniques to improve the Sentinel System's capacity to identify health outcomes of interest for post-market safety assessments [95].

However, because of the desire to use unstructured text regularly in the activities of DDNs, there is also a strong interest among DDNs to find ways of formally incorporating unstructured text into the CDM. One approach is to store the entire raw text directly as a single field in the CDM (e.g., as a character string). This approach is simple to implement, has high fidelity (i.e., minimizes information loss), and offers the most flexibility in how ML may be applied to the unstructured text in different activities (e.g., end-to-end vs pipeline approaches). However, it can also create unwieldy storage requirements, and for DDNs that normally allow for sharing of de-identified individual-level data, the highly sensitive nature of information stored in clinical text cannot be easily masked, thus creating significant privacy and confidentiality concerns for both patients and institutions [96]. In addition, the raw text cannot be immediately analyzed and must still be further processed by NLP tools before it can be used in downstream activities or combined with structured data. Another approach for incorporating unstructured text into the CDM involves processing the information upfront, which typically requires using NLP tools to extract information of interest from the raw clinical text (e.g., mentions and attributes of clinical concepts) and then encoding the output as structured data within the CDM (i.e., as a set of NLP-derived fields). The clear advantage of this approach is that the raw text is stored as structured data in the CDM, which can be immediately used and more easily de-identified. However, this approach requires significantly more time and resources to implement, including a data management team with expertise in NLP methods, and is more susceptible to information loss or even misclassification depending on the quality of the NLP tools used. In addition, this approach requires foreseeing the eventual use cases of unstructured text in future projects to ensure that the appropriate information and insights have been extracted at analysis time.

These two approaches for incorporating unstructured text into the CDM are not mutually exclusive, and one or both may be used. For example, the OMOP CDM includes two tables for unstructured text, one of which can store the original content of a note and another that can store the encoded output after applying NLP, where each row in the table represents a single extracted term from a note [97]. To facilitate the use of open-source NLP tools on unstructured text stored in the OMOP CDM, the OHDSI NLP working group has created several wrappers for NLP tools such as cTAKES and MetaMap that can be used by the OHDSI community [98]. Other commercial NLP tools also exist that can be used with raw text stored in the OMOP CDM. For example, a Health Insurance Portability and Accountability Act-compliant NLP service recently launched under Amazon Web Services, called Amazon Comprehend Medical [99], uses state-of-the-art deep learning models to extract clinical mentions and insights from unstructured text in the OMOP CDM and then writes the extracted insights back into the OMOP CDM using standardized ontological codes [100]. Although other DDNs such as the Sentinel System and PCORnet have not yet expanded their CDMs to include fields for unstructured text, they are actively pursuing efforts in this area and have funded several ongoing projects to explore scalable NLP processes for extracting clinical features from unstructured text and optimal approaches for incorporating these extracted insights into their CDMs to support future activities [101–103].

4.4 Scenario 4: Less Granular Data Shared

In this last scenario, we consider a DDN with the same characteristics as the base case, but rather than its data partners sharing individual-level data with the analysis center, the data partners share only summary-level data. Unlike Scenarios 2 and 3, where moving rightward along the spectrum created additional challenges for the feature engineering process, moving rightward along the spectrum in this scenario creates additional challenges for the model fitting process. These challenges arise because although the DDN in theory has access to data from multiple data partners, the possible analytic options are constrained by the inability of its data partners to share individual-level data.

One simple approach for applying ML under these constraints is to have each site fit its own ML model. These site-specific models can be compared and contrasted, or in the case of causal inference activities, the final causal parameter estimates can be pooled via a meta-analysis. With this approach, the capacity to properly compare and contrast findings across sites is greatly facilitated by using standardized processes and common programs with the CDM. These measures allow for the ML analyses to be conducted in a timely, consistent, transparent, and reproducible manner across multiple sites. For example, the OHDSI community has developed a standardized analytics pipeline [13] to guide its collaborators in developing and validating individual-level prediction models while making efforts to follow best practices [104] and limit potential causes of bias (e.g., by validating phenotypes, assessing data quality, and clearly specifying the target population). The entire analytics pipeline, from problem design to reliable model development and evaluation, can be implemented using open-source tools and packages developed by OHDSI collaborators to facilitate its timely and consistent implementation.

As a variation on this approach, DDNs may train a ML model in one site and apply the final model in another site. In addition to being potentially more efficient (e.g., when a manual chart review is required to determine the phenotype status), this modified approach may be preferred when the sample size at some sites is too small or when one site has a substantially larger sample size than the others. This approach also allows for the evaluation of a model's external validity and transportability across sites (i.e., generalizability)—an important task that is greatly facilitated when both sites use a CDM and standardized programs [105]. For example, in an OHDSI study where ML was used to develop a prognostic model for hemorrhagic transformation

[106], the final model (developed in one US database) was externally validated in ten databases from three different continents and found to be fairly transportable (Table 2). However, this approach may not be suitable if the model outcome or important features in the model are not available at the new site [104]. In cases where the model outcome is not available at the new site, the ML model may instead be used to impute outcomes at the new site, where the validity of the imputed outcomes can be indirectly assessed by comparing the distribution of imputed outcomes with known population-level statistics. For example, in another OHDSI study using ML to predict cause of death [107], the final model was not only externally validated (using one of the databases containing cause-of-death information), but also used to impute cause of death in three databases where this information was not available (Table 2). Finally, as a further extension of approaches involving site-specific ML models, DDNs may even choose to have multiple sites each develop a ML model and then externally validate the site-specific models in each of the other sites, a rotating model development-validation procedure coined "iterative pairwise external validation" [108]. EHDEN collaborators recently used this approach to develop and externally validate a collection of site-specific prognostic models to predict the 1-year risk of heart failure in patients starting a second pharmacotherapy for the treatment of type 2 diabetes mellitus [108]. Across five database-specific models that were developed, three models were found to consistently achieve comparable performance when externally validated across the remaining four databases (Table 2).

The aforementioned approaches, however, still only use data from a single site to train a ML model and therefore do not harness the full potential of DDNs to use data from multiple sites to develop more generalizable and robust ML models. In contrast, federated learning is an approach that allows multiple sites to collaboratively train a global ML model in a decentralized manner, such that sites can learn together but without sharing their private individual-level data [79]. Typically, this decentralized learning process occurs iteratively over multiple rounds, where at the start of each training round, sites start with the current version of the global model and use their local data to perform further training, sharing only summary-level characteristics or updates (e.g., parameters or gradients) from their locally trained models with a central coordinating server or other sites in the network [96]. These local updates are then aggregated and used to revise the global model, which is returned to sites for further training until the global model reaches the pre-specified convergence criteria. In this way, federated learning brings the model to the data, rather than the data to the model, and allows sites to protect their sensitive data while still collaborate to build more accurate and robust ML models [96]. Though appealing, this approach is challenging to implement in practice for several reasons. First, coordinating the logistics of such a distributed learning protocol, including the back-and-forth exchange of information during training iterations, can be a burdensome task. In addition, selecting the strategy with which to aggregate the model updates across sites requires careful attention and consideration, especially in the presence of heterogeneous (i.e., non-independent and non-identically distributed) data distributions, where a simple averaging of models across sites may not perform well [109, 110]. Finally, although sharing model parameters and gradients is more privacy protecting than sharing de-identified individual-level data (as in a centralized approach), it can still pose privacy risks owing to the amount of information that extremely flexible ML algorithms, such as deep learning models, can "memorize" about the training data ("information leakage") [96]. Additional measures such as differential privacy or learning from encrypted data can be used to further reduce the risk of information leakage, but they can also increase communication costs, training time, and reduce model performance [96, 111]. Such countermeasures may be neither desirable nor needed in the presence of proper governance and trusted collaborative relationships between data partners in a DDN. Despite these challenges, recent studies using federated learning with different ML architectures have shown that it is possible to achieve levels of performance comparable to models trained using a centralized approach [112] and better than locally trained models. For example, in a multisite study [113] using structured EHR data (e.g., laboratory data and vital signs) and chest X-ray images to predict the future oxygen requirements of symptomatic patients with COVID-19, the global federated deep learning model, trained using data from 20 clinical sites around the world, outperformed all local models that were trained at a single site using that site's data (Table 2). Federated learning is still very much an emerging and active area of research that will continue to develop in the years to come [96], not only for predicting health outcomes, but also for estimating causal effects, which to our knowledge has not yet been extensively explored for doubly-robust causal estimation frameworks such as targeted maximum likelihood estimation [114] that are becoming increasingly popular in pharmacoepidemiologic research.

5 Conclusions

Many opportunities exist for DDNs to use ML in pharmacoepidemiologic and pharmacovigilance studies. From phenotyping activities to signal detection, causal inference, and forecasting activities, the use of data-adaptive ML methods offers the potential to more fully capitalize on the larger amounts of electronic health data made accessible through the formation of DDNs.

However, the siloed storage and diverse nature of databases in DDNs also create unique challenges and considerations when applying ML. Many of these challenges stem from a DDN's practical data-related characteristics in terms of the modality of source data used, the extent to which sites within a DDN standardize their source data, and the level of granularity with which sites in a DDN share their data. In this paper, we presented several scenarios of DDNs with different characteristics (i.e., locations along the spectrum) of these data-related factors, and in each case, discussed challenges that DDNs may face when during the ML process. We also discussed possible approaches for addressing these challenges, including issues to consider and efforts that real-world DDNs have taken or are currently taking to help mitigate these challenges.

Ultimately, how a DDN chooses to implement ML and address challenges (e.g., whether and how to consider information outside the CDM, whether to fit site-specific models or use a federated learning approach) will involve a balancing act across three constraints: performance, price, and privacy. Ideally, DDNs would be able to develop the most accurate, robust, and generalizable models (or in the case of causal inference activities, obtain the most precise and consistent estimates of causal parameters) for a reasonable price (in terms of cost, resources, and effort involved) while maintaining adequate privacy of the data at all sites. In practice, however, achieving this ideal is often impossible. Thus, choosing amongst different approaches for performing ML in DDNs usually involves making a trade-off between these constraints. These decisions may also be further influenced by additional considerations such as the degree of suspected heterogeneity between databases, number of observations at each site, and the objectives of the task at hand. For example, in phenotyping or forecasting activities where only summary-level data can be shared and the dataset at each site is sufficiently large, use of a federated learning approach may not yield significant gains over simpler approaches such as fitting site-specific models, and thus the additional cost and effort required to implement a federated learning approach may not be warranted. However, federated learning may be more justifiable when there are many sites, each with only a small number of observations that alone cannot support the training of more complex ML algorithms, such as deep learning models, that often require a large amount of training data.

The use of ML in DDNs also creates greater opportunities or the impetus to address certain issues that often plague the use of ML in single-database settings. For example, while ML models developed in single-site settings are rarely and slowly (e.g., over years) externally validated, the collaboration of multiple data partners using a CDM and standardized programs and code in DDNs greatly facilitates the ability to easily and quickly (e.g., in months) externally validate ML models and identify models with greater generalizability [105]. Machine learning models developed in single-site settings also often suffer from a lack of transparency in the model development process; in DDNs, however, issues of transparency and reproducibility must be directly addressed because of the need for multiple data partners to work together and process a common set of ML models. Finally, as researchers may observe unusual or heterogeneous outputs from ML models across data partners in a DDN, the impetus to interpret ML models and explain their outputs may be encountered more frequently in DDNs than in single-site settings, as data partners work to recognize and rectify potential (unwanted) sources of heterogeneity, such as data errors and idiosyncrasies in coding and documentation practices across sites—a task that undoubtedly becomes more challenging with the use of more complex ML approaches. Taken together, by facilitating the development of more robust, generalizable, reproducible, and interpretable ML models, the use of ML in DDNs may increase the likelihood that the resulting models, and their outputs, can effectively enable or enhance decision making by clinicians, healthcare institutions, and regulatory bodies.

In conclusion, there is great potential, and great desire, to use ML in DDNs to enhance their activities in pharmacoepidemiology and pharmacovigilance. Indeed, the time is ripe for this emerging area of interest among DDNs because of not only the recent methodologic advancements in the field, but also the valuable groundwork that many DDNs have already laid through investments in harmonizing datasets to a CDM, developing standardized processes, tools, and analytics, and building collaborative relationships between data partners and with various stakeholders. The future holds much promise for the use of ML in DDNs, and we expect that the utility of these data-adaptive methods for enhancing pharmacoepidemiologic and pharmacovigilance studies will likely continue to increase in the years to come.

Declarations

Funding JW and ST are funded in part by a grant from the Agency for Healthcare Research and Quality (R01HS26214). DPA and PRR receive funding from The European Health Data and Evidence Network, which receives funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 806968; the Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme and the European Federation of Pharmaceutical Industries and Associations.

Conflicts of Interest/Competing Interests JW is a consultant to Hoffmann-La Roche Limited for unrelated work. DPA has received grant support from Amgen, Chesi-Taylor, Novartis, and UCB Biopharma; his department has received advisory or consultancy fees from Amgen, Astellas, AstraZeneca, Johnson and Johnson, and UCB Biopharma and fees for speaker services from Amgen and UCB Biopharma; Janssen, on behalf of IMI-funded EHDEN and EMIF consortiums, and Synapse Management Partners have supported training programs organized by DPA's department and open for external participants. PRR works for a research group that has received unconditional research grants from Boehringer-Ingelheim, GSK, Janssen Research & Development, Novartis, Pfizer, Yamanouchi, and Servier. RJD has served as a principal investigator for grants from Bayer, Vertex, and Novartis to Brigham and Women's Hospital. JMR is an employee of Janssen R&D and is a shareholder of Johnson and Johnson. ST is a consultant to Pfizer, Inc. and Merck & Co., Inc. for unrelated work.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Availability of Data and Material Not applicable.

Code Availability Not applicable.

Authors' Contributions JW and ST conceptualized and drafted the article. DPA, PRR, RJD, JMR, and ST critically reviewed the manuscript and provided important interpretation and critique of intellectual content. All authors have given final approval of the version to be published and agree to be accountable for all aspects of the work.

References

1. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform. 2016;2:S48-61.
2. Murray MD. Use of data from electronic health records for pharmacoepidemiology. Curr Epidemiol Rep. 2014;1:186–93.
3. Toh S, Pratt N, Klungel O, Gagne JJ, Platt RW. Distributed networks of databases analyzed using common protocols and/or common data models. Pharmacoepidemiology. John Wiley & Sons, Ltd; 2019: p. 617–38. Available from: https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1002/9781119413431.ch25. [Accessed 1 Aug 2021].
4. Burn E, You SC, Sena A, Kostka K, Abedtash H, Abrahao MTF, 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. MedRxiv Prepr Serv Health Sci. 2020;2020.04.22.20074336.
5. 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:S45-51.
6. US Food and Drug Administration. Sentinel system five-year strategy, 2019-2023. 2019 Jan. Available from: https://www.fda.gov/media/120333/download. [Accessed 20 Feb 2022].
7. Platt RW, Henry DA, Suissa S. The Canadian Network for Observational Drug Effect Studies (CNODES): reflections on the first eight years, and a look to the future. Pharmacoepidemiol Drug Saf. 2020;29(Suppl. 1):103–7.
8. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. Npj Digit Med. 2021;4:1–23.
9. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2018;22:1589–604.
10. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019;7:e12239.
11. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. Npj Digit Med. 2021;4:1–13.
12. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. Pac Symp Biocomput Pac Symp Biocomput. 2020;25:295–306.
13. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed. 2021;211:106394.
14. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A review of automatic phenotyping approaches using electronic health records. Electronics. 2019;8:1235.
15. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional Propensity Score Algorithm? Epidemiology. 2018;29:191–8.
16. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol. 2018;47:2005–14.
17. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319:1317–8.
18. AsPEN collaborators, Andersen M, Bergman U, Choi N-K, Gerhard T, Huang C, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. Pharmacoepidemiol Drug Saf. 2013;22:700–4.
19. Suissa S, Henry D, Caetano P, Dormuth CR, Ernst P, Hemmelgarn B, et al. CNODES: the Canadian network for observational drug effect studies. Open Med. 2012;6:e134–40.
20. European Health Data Evidence Network. Available from: https://www.ehden.eu/. [Accessed 7 Sep 2021].
21. 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.
22. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol. 2021;129:60–7.
23. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–8.
24. 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.
25. McNeil MM, Gee J, Weintraub ES, Belongia EA, Lee GM, Glanz JM, et al. The Vaccine Safety Datalink: successes and challenges monitoring vaccine safety. Vaccine. 2014;32:5390–8.
26. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med. 2016;71:57–61.
27. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci. 2018;1:53–68.

28. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. Curr Epidemiol Rep. 2018;5:331–42.
29. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for patient phenotyping from electronic health records. Sci Rep. 2020;10:1432.
30. Ong M-S, Klann JG, Lin KJ, Maron BA, Murphy SN, Natter MD, et al. Claims-based algorithms for identifying patients with pulmonary hypertension: a comparison of decision rules and machine-learning approaches. J Am Heart Assoc. 2020;9:e016648.
31. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23:1007–15.
32. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350:h1885.
33. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. J Biomed Inform. 2019;97:103258.
34. Bate A, Hornbuckle K, Juhaeri J, Motsko SP, Reynolds RF. Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance. Ther Adv Drug Saf. 2019;10:2042098619864744.
35. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. 2013;22:39–56.
36. Arnaud M, Bégaud B, Thurin N, Moore N, Pariente A, Salvo F. Methods for safety signal detection in healthcare databases: a literature review. Expert Opin Drug Saf. 2017;16:721–32.
37. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54:315–21.
38. Wang SV, Maro JC, Baro E, Izem R, Dashevsky I, Rogers JR, et al. Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. Epidemiology. 2018;29:895–903.
39. Wang SV, Maro JC, Gagne JJ, Patorno E, Kattinakere S, Stojanovic D, et al. A general propensity score for signal identification using tree-based scan statistics. Am J Epidemiol. 2021;190:1424–33.
40. Reps JM, Garibaldi JM, Aickelin U, Gibson JE, Hubbard RB. A supervised adverse drug reaction signalling framework imitating Bradford Hill's causality considerations. J Biomed Inform. 2015;56:356–68.
41. Liu F, Jagannatha A, Yu H. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. Drug Saf. 2019;42:95–7.
42. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020;27:3–12.
43. Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository. AMIA Annu Symp Proc. 2011;2011:1270–9.
44. Hohl CM, Kuramoto L, Yu E, Rogula B, Stausberg J, Sobolev B. Evaluating adverse drug event reporting in administrative data from emergency departments: a validation study. BMC Health Serv Res. 2013;13:473.
45. Kennedy EH. Semiparametric theory and empirical processes in causal inference. In: He H, Wu P, Chen D-G (Din), editors. Statistical causal inferences and their application in public health research. Cham: Springer International Publishing; 2016: p. 141–67. doi: https://doi.org/10.1007/978-3-319-41259-7_8.
46. Schneeweiss S, Suissa S. Advanced approaches to controlling confounding in pharmacoepidemiologic studies. Pharmacoepidemiology. John Wiley & Sons, Ltd; 2019: p. 1078–107. doi: https://doi.org/10.1002/9781119413431.ch43.
47. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. Epidemiology. 2021;32:393–401.
48. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163:1149–56.
49. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. J Clin Epidemiol. 2010;63:826–33.
50. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:25.
51. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20:512–22.
52. Ju C, Combs M, Lendle SD, Franklin JM, Wyss R, Schneeweiss S, et al. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. J Appl Stat. 2019;46:2216–36.
53. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM. Using super learner prediction modeling to improve high-dimensional propensity score estimation. Epidemiology. 2018;29:96–106.
54. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol. 2018;10:771–88.
55. Soyiri IN, Reidpath DD. An overview of health forecasting. Environ Health Prev Med. 2013;18:1–9.
56. Wright G, Lawrence MJ, Collopy F. The role and validity of judgment in forecasting. Int J Forecast. 1996;12:1–8.
57. Farinholt P, Park M, Guo Y, Bruera E, Hui D. A comparison of the accuracy of clinician prediction of survival versus the Palliative Prognostic Index. J Pain Symptom Manag. 2018;55:792–7.
58. Saposnik G, Cote R, Mamdani M, Raptis S, Thorpe KE, Fang J, et al. JURaSSiC. Neurology. 2013;81:448–55.
59. Rojas JC, Lyons PG, Jiang T, Kilaru M, McCauley L, Picart J, et al. Accuracy of clinicians' ability to predict the need for intensive care unit readmission. Ann Am Thorac Soc. 2020;17:847–53.
60. Sun JW, Franklin JM, Rough K, Desai RJ, Hernández-Díaz S, Huybrechts KF, et al. Predicting overdose among individuals prescribed opioids using routinely collected healthcare utilization data. PLoS ONE. 2020;15:e0241083.
61. Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among Medicare beneficiaries with opioid prescriptions. JAMA Netw Open. 2019;2:e190968.
62. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. JAMA Netw Open. 2019;2:e1915997.
63. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am J Psychiatry. 2018;175:951–60.
64. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. Npj Digit Med. 2020;3:1–9.
65. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data

in analysis of electronic health records. WIREs Comput Stat. 2022;2:e1549.

66. Kong H-J. Managing unstructured big data in healthcare system. Healthc Inform Res. 2019;25:1–2.
67. Speech and language processing. Available from: https://web.stanford.edu/~jurafsky/slp3/. [Accessed 13 Sep 2021].
68. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. J Biomed Inform. 2020;109:103526.
69. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34–49.
70. OHDSI Natural Language Processing Working Group. https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg#objective. Accessed 13 Sept 2021.
71. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics. 2021;39:275–85.
72. The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/. Accessed 13 Sept 2021.
73. Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias. Pharmacoepidemiol Drug Saf. 2019;2:2.
74. Health Analytics Data-to-Evidence Suite (HADES). Observational Health Data Sciences and Informatics; 2021. Available from: https://github.com/OHDSI/Hades. [Accessed 15 Oct 2021].
75. Klungel OH, Kurz X, de Groot MCH, Schlienger RG, Tcherny-Lessenot S, Grimaldi L, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiol Drug Saf. 2016;25(Suppl. 1):156–65.
76. 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:S4-10.
77. Jeon G, Sangaiah AK, Chen Y-S, Paul A. Special issue on machine learning approaches and challenges of missing data in the era of big data. Int J Mach Learn Cybern. 2019;10:2589–91.
78. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC). 2016;4:1244.
79. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50:S21–9.
80. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013;51:S22–9.
81. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC). 2015;3:1052.
82. Adimadhyam S, Barreto EF, Cocoros NM, Toh S, Brown JS, Maro JC, et al. Leveraging the capabilities of the FDA's sentinel system to improve kidney care. J Am Soc Nephrol. 2020;31:2506–16.
83. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021;28:2251–7.
84. Shi X, Li X, Cai T. Spherical regression under mismatch corruption with application to automated knowledge translation. J Am Stat Assoc. 2020;2:1–12.
85. Sentinel Initiative. Using unsupervised learning to harmonize data across data systems. Available from: https://www.sentinelinitiative.org/methods-data-tools/methods/using-unsupervised-learning-harmonize-data-across-data-systems. [Accessed 19 Sep 2021].
86. Schuemie MJ, Madigan D, Ryan PB, Reich C, Suchard MA, Berlin JA, et al. Comment on "How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias." Pharmacoepidemiol Drug Saf. 2019;28:1032–3.
87. Arterburn D, Wellman R, Emiliano A, Smith SR, Odegaard AO, Murali S, et al. Comparative effectiveness and safety of bariatric procedures for weight loss: a PCORnet cohort study. Ann Intern Med. 2018;169:741–50.
88. Hurst JH, Liu Y, Maxson PJ, Permar SR, Boulware LE, Goldstein BA. Development of an electronic health records datamart to support clinical and population health research. J Clin Transl Sci. 2020;5:e13.
89. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO research network virtual data warehouse: a public data model to support collaboration. EGEMS. 2014;2:1049.
90. PCORnet® COVID-19 common data model launched, enabling rapid capture of insights on patients infected with the novel coronavirus. National Patient-Centered Clinical Research Network. 2020. Available from: https://pcornet.org/news/pcornet-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/. [Accessed 20 Sep 2021].
91. Cocoros NM, Fuller CC, Adimadhyam S, Ball R, Brown JS, Pan GJD, et al. A COVID-19-ready public health surveillance system: the Food and drug administration's sentinel system. Pharmacoepidemiol Drug Saf. 2021;30:827–37.
92. Sentinel Initiative. Validation of anaphylaxis using machine learning. Available from: https://www.sentinelinitiative.org/methods-data-tools/methods/validation-anaphylaxis-using-machine-learning. [Accessed 29 Sep 2021].
93. Sentinel Initiative. Validation of acute pancreatitis using machine learning and multi-site adaptation for anaphylaxis. Available from: https://www.sentinelinitiative.org/methods-data-tools/methods/validation-acute-pancreatitis-using-machine-learning-and-multi-site. [Accessed 21 Sep 2021].
94. Carrell DS. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. 2021. Available from: https://sentinelinitiative.org/sites/default/files/documents/ICPE%20Presentation%20-%20Improving%20Methods%20of%20Identifying%20Anaphylaxis%20for%20Medical%20Product%20Safety.pdf. [Accessed 20 Sep 2021].
95. Sentinel Initiative. Extending machine learning methods development in Sentinel: follow-up analyses for anaphylaxis algorithm and formalization of a general phenotyping framework (phase 3). Available from: https://www.sentinelinitiative.org/methods-data-tools/methods/extending-machine-learning-methods-development-sentinel-follow-analyses. [Accessed 29 Sep 2021].
96. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. Npj Digit Med. 2020;3:1–7.
97. GitHub. NOTE NLP table · Issue #85 · OHDSI/CommonDataModel. Available from: https://github.com/OHDSI/CommonDataModel/issues/85. [Accessed 4 Sep 2021].
98. GitHub. OHDSI/NLPTools. NLPTools/Wrappers at master. Available from: https://github.com/OHDSI/NLPTools. [Accessed 21 Sep 2021].
99. Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend Medical: a named entity recognition and relationship extraction web xervice. ArXiv191007419 Cs. 2019. Available from: http://arxiv.org/abs/1910.07419. [Accessed 6 Sep 2021].

100. AWS. Map clinical notes to the OMOP Common Data Model and healthcare ontologies using Amazon Comprehend Medical. 2019. Available from: https://aws.amazon.com/blogs/machine-learning/map-clinical-notes-to-the-omop-common-data-model-and-healthcare-ontologies-using-amazon-comprehend-medical/. [Accessed 21 Sep 2021].
101. Sentinel Initiative. Representation of unstructured data across common data models. Available from: https://www.sentinelinitiative.org/methods-data-tools/methods/representation-unstructured-data-across-common-data-models. [Accessed 21 Sep 2021].
102. Sentinel Initiative. Advancing scalable natural language processing approaches for unstructured electronic health record data. Available from: https://www.sentinelinitiative.org/methods-data-tools/methods/advancing-scalable-natural-language-processing-approaches-unstructured. [Accessed 21 Sep 2021].
103. PCORI (Patient-Centered Outcomes Research Institute). Improving methods for identifying social, behavioral, and clinical factors in doctors' notes in electronic health records. 2019. Available from: https://www.pcori.org/research-results/2019/improving-methods-identifying-social-behavioral-and-clinical-factors-doctors%E2%80%99. [Accessed 21 Sep 2021].
104. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25:969–75.
105. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20:102.
106. Wang Q, Reps JM, Kostka KF, Ryan PB, Zou Y, Voss EA, 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.
107. Kim C, You SC, Reps JM, Cheong JY, Park RW. Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. J Am Med Inform Assoc. 2021;28:1098–107.
108. Williams RD, Reps JM, Kors JA, Ryan PB, Steyerberg E, Verhamme KM, et al. Using iterative pairwise external validation to contextualize prediction model performance: a use case predicting 1-year heart-failure risk in diabetes patients across five data sources. Drug Saf. 2022;2:2.
109. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-IID data. ArXiv180600582 Cs Stat. 2018. Available from: http://arxiv.org/abs/1806.00582. [Accessed 29 Sep 2021].
110. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. ArXiv181206127 Cs Stat. 2020. Available from: http://arxiv.org/abs/1812.06127. [Accessed 27 Sep 2021].
111. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37:50–60.
112. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. Npj Digit Med. 2021;4:1–8.
113. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;2:1–9.
114. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat. 2006. https://doi.org/10.2202/1557-4679.1043/html.