

1 Applying Machine Learning in Distributed Data Networks for
2 Pharmacoepidemiologic and Pharmacovigilance Studies: Opportunities,
3 Challenges, and Considerations
4

5 Running title: Applying Machine Learning in Distributed Data Networks

6 Jenna Wong¹, Daniel Prieto-Alhambra^{2,3}, Peter R. Rijnbeek³, Rishi J. Desai⁴, Jenna M. Reps⁵, Sengwee

7 Toh¹

8

9 ¹Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute,
10 Boston, MA

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

13 ³Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands

14 ⁴Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard
15 Medical School, Boston, MA

16 ⁵Janssen Research & Development, LLC, Titusville, NJ

17

18 Corresponding author

19 Sengwee Toh, ScD

20 Division of Therapeutics Research and Infectious Disease Epidemiology (TIDE)

21 Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute

22 401 Park Drive, Suite 401 East

23 Boston, MA 02215

24 darren_toh@harvardpilgrim.org

25

26 Key Points

- 27
- Many opportunities exist for distributed data networks (DDNs) to use machine learning in
28 pharmacoepidemiologic and pharmacovigilance studies; however, the practical data-related
29 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
30 and present different approaches for addressing these challenges, including issues for
31 consideration and examples of how real-world DDNs have addressed or are working to help
32 mitigate these challenges.
 - The use of machine learning in DDNs is an emerging area of interest that holds much promise,
33 and the utility of these data-adaptive modeling methods for enhancing pharmacoepidemiologic
34 and pharmacovigilance studies will likely continue to increase in the years to come.
35
36

37

38 Abstract

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

57

58

59

60 1 Introduction

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

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

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

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

103 Throughout our discussion, we use the term “machine learning” to refer to highly flexible and data-
104 adaptive algorithms that can automatically learn complex associations in high-dimensional data with
105 minimal human guidance (e.g., random forests, support vector machines with gaussian kernels, deep

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

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 & Evidence Network (EHDEN) [20]
- Health Care Systems Research Network (HCSRN) [21]
- National Patient-Centered Clinical Research Network (PCORnet®) [22]
- Observational Health Data Science and Informatics (OHDSI) Collaborative [23]
- Sentinel System [24]
- Vaccine Safety Datalink [25]

112

113 [2 Opportunities for Machine Learning](#)

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

121 2.1 Computable Phenotyping

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

128 Phenotyping algorithms have traditionally consisted of expert-defined rules based on structured health
129 data, such as medical codes or laboratory tests [26,27]. Although interpretable and relatively simple to
130 implement, rule-based phenotyping algorithms can only accommodate pre-existing knowledge or beliefs
131 about a medical condition [26] and may be particularly challenging to develop when complex clinical
132 criteria and tacit knowledge are required to make a diagnosis [28]. Furthermore, limiting phenotyping
133 algorithms to using only structured data forgoes the opportunity to capitalize on the abundance of
134 clinical information stored in unstructured data (e.g., clinical text) – above and beyond that available in
135 structured data, which are more challenging to extract but could be valuable for phenotyping activities
136 [27].

137 In contrast, the use of ML for disease classification represents a more data-driven approach that can
138 consider a multitude of clinical features to identify latent associations and potentially new phenotyping
139 definitions. This process typically involves representing the information to be considered in the
140 phenotyping algorithm as a feature vector, tagging a set of observations with labels (i.e., having or not
141 having the phenotype of interest), and then allowing the data to train a supervised ML algorithm that
142 maps the input features to the labels [9,26]. ML may also be used to facilitate the extraction of
143 potentially relevant phenotypic information from clinical text as part of a natural language processing

144 (NLP) tool – the output of which can then be used to identify phenotypes directly [29] or create NLP-
145 derived features to be combined with other structured data in a downstream ML algorithm [27].
146 Phenotyping algorithms have been found to achieve better performance when developed using
147 supervised ML compared to conventional decision rules [30] and when presented with features derived
148 from both structured and unstructured data compared to structured data alone [31,32]. In addition, ML
149 has been used in other unique ways to enhance phenotyping activities, such as to estimate “probabilistic
150 gold standard” phenotype probabilities on large groups of individuals to facilitate the efficient
151 estimation of complete validation parameters (i.e., sensitivity and specificity, in addition to predictive
152 values) for simple rule-based phenotyping algorithms commonly used for cohort development (e.g., 1 or
153 more occurrences of a diagnosis code for the phenotype) [33].

154 2.2 Safety Signal Detection

155 Signal detection activities monitor the safety of medical products in real-world settings by detecting the
156 emergence of new or unsuspected adverse events that may be associated with a product’s use, where
157 the identified signals are then further investigated for evidence of a potential causal association
158 (discussed in Section 2.3). Compared to the other domains of activities discussed in this section, safety
159 signal detection activities are generally more specific to the field of pharmacoepidemiology and
160 pharmacovigilance. Though signal detection activities have been traditionally performed using data from
161 spontaneous reporting systems (SRS), electronic health databases are being increasingly recognized as a
162 valuable alternative data source due to their potential to address many limitations of SRS data (e.g.,
163 reporting bias, inability to estimate rates due to lack of population denominator), among other reasons
164 [34].

165 A variety of approaches for signal detection have been tested with electronic health databases, including
166 approaches that transport methods originally designed for SRS data (e.g., disproportionality analyses),

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

184 2.3 Causal Inference

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

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

204 Parametric regression models are commonly used to estimate nuisance functions, but the strong
205 functional assumptions they make are prone to misspecification, particularly when there are many
206 covariates with potentially complex associations. Alternatively, data-adaptive supervised ML techniques
207 offer a more flexible modeling approach for estimating nuisance functions that impose less restrictive
208 assumptions on covariates, thus increasing the likelihood of attaining properly specified nuisance
209 models for more valid inferences [15,49], especially when used with cross-fitting estimation procedures
210 [47]. Nuisance functions estimated using ensemble ML techniques [50], which can consider collections
211 of different ML algorithms and different covariate sets (e.g., using various thresholds of top-ranked
212 covariates from the high-dimensional propensity score algorithm [51]), have been shown to yield more
213 consistent estimates of causal parameters [47,52,53]. ML methods may also be useful for further
214 automating the confounding adjustment process by efficiently extracting and prioritizing appropriate

215 covariates from high-dimensional spaces in claims data, structured EHR data, and unstructured EHR data
216 (e.g., word stems or N-grams) for inclusion in nuisance function models [54].

217 2.4 Forecasting

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

228 In contrast, statistical forecasting uses statistical models based on historical data to predict the
229 occurrence of future health events (i.e., for prognostic modeling). Such prognostic models not only
230 enable population-level forecasting, but they can also be used to further inform clinicians' judgements
231 and enhance decision-making at the patient level (e.g., as decision support tools). Like with the use of
232 supervised ML for diagnostic modeling in phenotyping activities, the use of supervised ML for prognostic
233 modeling offers a data-adaptive approach for identifying potentially complex associations and
234 interactions between a plethora of features to predict future health outcomes. Supervised ML has been
235 used with electronic health data to create accurate risk prediction models for many health events for
236 which prospective surveillance, prevention, early intervention, and advanced planning is invaluable,

237 including opioid overdose [60,61], cancer-related mortality [62], suicidality [63], and high healthcare
238 costs [64].

239 3 Practical Data-Related Factors

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

246 3.1 Modality of Source Data

247 Electronic health data exist in formats that are either structured or unstructured [65]. Structured data
248 have a well-defined format (e.g., tables), where information is often stored using standardized values
249 and can be easily extracted and analyzed. Information commonly stored as structured data include
250 patient demographics (e.g., age), medications, coded diagnoses and procedures, certain laboratory tests
251 (e.g., international normalized ratio), and certain quantitative clinical measures (e.g., blood pressure). In
252 contrast, unstructured data (e.g., clinical notes, discharge summaries, radiology and pathology reports,
253 and medical images) contain valuable information beyond that captured in structured data, but their
254 lack of any pre-defined structure makes them challenging for computers to process and analyze [65].
255 Depending on the source of electronic health data, one or both data modalities may exist. For example,
256 administrative claims databases contain exclusively structured data, while EHR databases typically
257 contain a mixture of structured and unstructured data, with the majority of these data often being
258 unstructured [66].

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

277 3.2 Degree of Data Standardization

278 In their native environments, most electronic health databases vary greatly in their schemas, content,
279 and coding terminologies [71]. Given the significant discrepancies that often exist between disparate
280 databases in a DDN, a range of approaches may be used to deal with these differences when conducting
281 a distributed analysis. At one end of the spectrum, DDNs can have all their data partners standardize the
282 format of their source data to a common data model (CDM). A CDM specifies a standardized structure

283 and set of tables to which data partners in a DDN convert their source data [3,71]. Some CDMs, such as
284 the Observational Medical Outcomes Partnership (OMOP) CDM – maintained by the OHDSI group and
285 also used by EHDEN and AsPEN – additionally standardize coding terminologies to a common vocabulary
286 [72], whereas other CDMs like the Sentinel CDM only store the original coded values to be later mapped
287 to a common vocabulary on a study-specific basis [73]. Although CDMs require significant up-front
288 investment to both develop the infrastructure and implement the specifications, the invested time and
289 effort pay increasing dividends as more studies are performed [3]. In particular, use of a CDM offers the
290 ability to apply validated software and tools that have been developed for the CDM, thus promoting not
291 only the rapid, reliable, and reproducible implementation of analyses across sites, but also reducing the
292 amount of site-specific programming required and coding errors encountered [71]. The Sentinel System,
293 PCORnet, and OHDSI are examples of DDNs that have developed libraries of customized tools for use
294 with data formatted to their CDMs [3]. For example, members of the OHDSI community may utilize
295 open-source R packages available in the Health Analytics Data-to-Evidence Suite (HADES) [74] to
296 perform various study analyses on data converted to the OMOP CDM.

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

306 CDM may be developed for only some data partners within a DDN, or only a small fraction of
307 information within data partners' source systems may be converted to a CDM.

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

316 3.3 Granularity of Shared Data

317 Once all sites have prepared their analytic datasets containing the necessary study variables for all
318 eligible individuals in their databases, there are different ways in which these individual-level datasets
319 may be used to perform the activity at hand. At one end of the spectrum, each site in the network can
320 share its individual-level dataset with the analysis center, which may also be a data-contributing site.
321 Once the individual-level datasets have been pooled, the result is a single centralized dataset [3]. This
322 process is distinct from a centralized data approach because only the final study datasets – not the
323 underlying data sources from which they were curated – are shared, subject to each data partner having
324 approved the request and reviewed its final dataset before sharing. Although this approach requires the
325 most granular level of data sharing (and is thus the least privacy-protecting), it offers the most analytic
326 freedom, such that decisions about the methods used and how they are applied to the data can be
327 based solely on the scientific needs of the study [76]. For example, if the analysis center felt it was best
328 to fit a separate treatment nuisance model for each site but a global outcome nuisance model across all

329 sites, then this plan could be easily implemented. To minimize privacy and confidentiality concerns, the
330 individual-level datasets are typically de-identified (e.g., by removing personal identifiers), and if
331 desired, extra care can be taken to reduce their dimensionality (e.g., by combining individual variables
332 into summary measures that contain essentially the same information using fewer variables) [3]. In
333 addition, the ability to share de-identified individual-level data generally requires proper governance,
334 appropriate data use agreements, and established collaborative relationships between sites, leading to a
335 shared sense of trust. Such elements are important pre-requisites that are necessary – though not
336 always sufficient – to allow for sharing of individual-level data between entities [3].

337 At the other end of spectrum, each site in the network can share only summary-level data. Different
338 types of summary-level data may be shared depending on the type of activity being performed,
339 methodologic approach being used, and degree of privacy protection desired. For example, when
340 conducting causal inference activities while sharing only summary-level data, a stratified analysis can be
341 implemented by having each site send aggregated counts of the total number of persons (or person-
342 time) and outcomes in each treatment group per stratum to the analysis center – essentially, a
343 coarsened version of the individual-level dataset [76]. Another approach is to use distributed regression,
344 which produces identical results to a centralized outcome regression analysis of individual-level data.
345 Distributed regression can be implemented by having each site fit a local regression on its own data and
346 share only intermediate model statistics (e.g., sums of squares and cross-products matrix) with the
347 analysis center, which then calculates the global parameter estimates, and if necessary, sends them
348 back to each site for additional processing to update the global parameters iteratively until a pre-
349 specified convergence criterion is met [3]. A third possible approach, which offers the greatest amount
350 of privacy protection, is to have each site conduct its own analysis and estimate a site-specific causal
351 parameter estimate that can then be pooled by the analysis center via meta-analysis [76]. Regardless of
352 the approach, when sites share only summary-level data, the analysis plan is typically more constrained

353 because it must consider not only the scientific needs of the study, but also the practical challenges of
354 being unable to combine and process individual-level data from each site. Besides sharing only
355 individual-level or summary-level data, other data sharing combinations are also possible. For example,
356 only some sites within a DDN may share individual-level data with the analysis center, or all sites may
357 share individual-level data for some projects but only summary-level data for other projects.

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

362 4 Challenges and Considerations

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

372 4.1 Scenario 1: Base Case

373 We first consider a DDN located at the leftmost end of the spectrum for all three data-related factors in
374 Figure 2. In this scenario, all sites in the DDN utilize only structured data from their source systems,
375 these structured data populate a CDM containing all information needed to create the inputs for an

376 activity, and all sites share their final, de-identified individual-level datasets with the analysis center. We
377 refer to this scenario as the “base case”, which also serves as the reference for scenarios 2-4.

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

387 From a scientific perspective, the base case still presents issues for consideration when applying ML due
388 to heterogeneity that may exist between databases in a DDN. In other words, although it is technically
389 possible to combine all the datasets and apply ML to the centralized dataset, *should* it be done?

390 Heterogeneity between databases can exist for a variety of reasons, including differences in data
391 encoding (e.g., data quality and coding practices) and content (e.g., available data elements and
392 domains), as well as variation in patient characteristics and care. To help disentangle true heterogeneity
393 in patient populations and care from data quality problems, it is important that DDNs conduct regular
394 and robust data quality assessments, ideally according to a systematic and conceptually-based
395 framework [78–81]. For example, to minimize data quality issues and errors that may arise during the
396 CDM creation process, the Sentinel System requires that all extracts from their data partners first pass
397 an extensive data quality review process [73,82]. The OHDSI collaborators have also developed the Data
398 Quality Dashboard [83] – an open-source tool that performs a series of systematic data quality checks
399 on databases mapped to the OMOP CDM to report potential data quality issues before these databases

400 are used in modeling activities [13]. To further identify and reduce potential heterogeneity in coding
401 practices between data partners, the Sentinel System is currently exploring the use of novel code
402 translation methods, which incorporate unsupervised learning and language translation methods, to
403 generate data-driven code mappings that could be used as a scalable and automated approach to help
404 address idiosyncrasies in the coding process and harmonize medical codes across databases [84,85].
405 Such initiatives highlight the importance of preserving the original data values to the greatest extent
406 possible when populating the CDM to maintain high fidelity and minimize information loss during the
407 conversion process – a guiding principle of both the Sentinel and OMOP CDMs (the OMOP CDM both
408 maps codes to a common vocabulary and retains the original source codes) [73,86].

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

423 potential influence of database heterogeneity on the study findings and select the most appropriate
424 approach for the task at hand.

425 4.2 Scenario 2: Less Standardized Data Available

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

432 When information of interest exists outside the CDM, the unstandardized data across sites creates
433 challenges during the feature engineering process. One approach for addressing this challenge is to
434 standardize the unstandardized information, which may be done by having sites create "data sidecars"
435 for the study (i.e., tables containing additional data elements that can be linked to the primary tables in
436 the CDM) [88] or by expanding the CDM itself to include new fields or tables. However, because these
437 approaches are often costly and time-consuming to implement, the anticipated benefits must be
438 carefully weighed against the anticipated costs. For example, the HCSRN requires that any specification
439 changes or table additions to its Virtual Data Warehouse (VDW) CDM are first proposed by workgroups,
440 then discussed by the VDW Implementation Group, and finally approved through a formal voting
441 process [89]. Often, these efforts are only pursued if the information of interest can be easily obtained
442 (low-effort), will be frequently used (high-yield), or is urgently required (high-demand). For example, to
443 support public health efforts during the COVID-19 pandemic, both PCORnet [90] and the Sentinel
444 System [91] added new fields to their core CDMs to capture important information needed to better
445 characterize infected patients. In addition, the Sentinel System, as part of its five-year strategy [6], is

446 looking to expand its CDM (currently mostly claims-based) to include additional fields containing more
447 granular clinical information from structured data fields in EHRs (e.g., vital signs and body mass index) –
448 an initiative that will broaden the range of stakeholder questions the Sentinel System can consider in
449 future projects.

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

459 4.3 Scenario 3: More Complex Data Modalities Used

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

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

480 However, because of the desire to use unstructured text regularly in the activities of DDNs, there is also
481 a strong interest among DDNs to find ways of formally incorporating unstructured text into the CDM.
482 One approach is to store the entire raw text directly as a single field in the CDM (e.g., as a character
483 string). This approach is simple to implement, has high fidelity (i.e., minimizes information loss), and
484 offers the most flexibility in how ML may be applied to the unstructured text in different activities (e.g.,
485 end-to-end versus pipeline approaches). However, it can also create unwieldy storage requirements,
486 and for DDNs that normally allow for sharing of de-identified individual-level data, the highly sensitive
487 nature of information stored in clinical text cannot be easily masked, thus creating significant privacy
488 and confidentiality concerns for both patients and institutions [96]. In addition, the raw text cannot be
489 immediately analyzed and thus still requires further processing by NLP tools before it can be used in
490 downstream activities or combined with structured data. Another approach for incorporating
491 unstructured text into the CDM is to do the information processing upfront. This process typically
492 involves using NLP tools to extract information of interest from the raw clinical text (e.g., mentions and

493 attributes of clinical concepts) and then encoding the output as structured data within the CDM (i.e., as
494 a set of NLP-derived fields). The clear advantage of this approach is that the raw text is stored as
495 structured data in the CDM, which can be immediately used and more easily de-identified. On the other
496 hand, this approach requires significantly more time and resources to implement, including a data
497 management team with expertise in NLP methods, and is more susceptible to information loss or even
498 misclassification depending on the quality of the NLP tools used. In addition, this approach requires
499 foreseeing the eventual use cases of unstructured text in future projects to ensure that the appropriate
500 information and insights have been extracted at analysis time.

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

517 4.4 Scenario 4: Less Granular Data Shared

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

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

538 As a variation on this approach, DDNs may train a ML model in one site and apply the final model in
539 another site. This approach has been used successfully in several OHDSI studies to develop ML models

540 for various health outcomes, such as hemorrhagic transformation after ischemic stroke [105] and cause
541 of death [106]. Besides being potentially more efficient (e.g., when manual chart review is required to
542 determine the phenotyping status), this modified approach may be preferred when the sample size at
543 some sites is too small or when one site has a substantially larger sample size than the others. This
544 approach also allows for the evaluation of a model’s external validity and transportability across sites
545 (i.e., generalizability) – an important task that is greatly facilitated when both sites use a CDM and
546 standard programs [107]. For example, in the OHDSI study where ML was used to develop a prognostic
547 model for hemorrhagic transformation [105], the final model (developed in one US database) was
548 externally validated in 10 databases from three different continents and found to be fairly transportable
549 (Table 2). However, this approach may not be suitable if the model outcome or important features in
550 the model are not available at the new site [104]. In cases where the model outcome is not available at
551 the new site, the ML model may instead be used to impute outcomes at the new site, where the validity
552 of the imputed outcomes can be indirectly assessed by comparing the distribution of imputed outcomes
553 with known population-level statistics. For example, in the OHDSI study using ML to predict cause of
554 death [106], the final model was not only externally validated (using one of the databases containing
555 cause-of-death information), but also used to impute cause of death in three databases where this
556 information was not available (Table 2). Finally, as a further extension of approaches involving site-
557 specific ML models, DDNs may even choose to have multiple sites each develop a ML model and then
558 externally validate the site-specific models in each of the other sites – a rotating model development-
559 validation procedure coined “iterative pairwise external validation” [108]. EHDEN collaborators recently
560 used this approach to develop and externally validate a collection of site-specific prognostic models to
561 predict the 1-year risk of heart failure in patients starting a second pharmacotherapy for treatment of
562 type 2 diabetes [108]. Across five databases-specific models that were developed, three models were

563 found to consistently achieve comparable performance when externally validated across the remaining
564 four databases (Table 2).

565 The aforementioned approaches, however, still only use data from a single site to train a ML model and
566 therefore do not harness the full potential of DDNs to use data from multiple sites to develop more
567 generalizable and robust ML models. In contrast, federated learning is an approach that allows multiple
568 sites to collaboratively train a global model in a decentralized fashion, such that sites can learn together
569 but without sharing their private, individual-level data [79]. Typically, this decentralized learning process
570 occurs iteratively over multiple rounds, where at the start of each training round, sites start with the
571 current version of the global model and use their local data to perform further training, sharing only
572 summary-level characteristics or updates (e.g., parameters or gradients) from their locally trained
573 models with a central coordinating server or other sites in the network [96]. These local updates are
574 then aggregated and used to revise the global model, which is returned to sites for further training until
575 the global model reaches the pre-specified convergence criteria. In this way, federated learning brings
576 the model to the data, rather than the data to the model, and allows sites to protect their sensitive data
577 while still collaborate to build more accurate and robust ML models [96]. Though appealing, this
578 approach is challenging to implement in practice for several reasons. First, coordinating the logistics of
579 such a distributed learning protocol, including the back-and-forth exchange of information during
580 training iterations, can be a burdensome task. In addition, selecting the strategy with which to aggregate
581 the model updates across sites requires careful attention and consideration, especially in the presence
582 of heterogeneous (i.e., non-identically distributed) data distributions, where a simple averaging of
583 models across sites may not perform well [109,110]. Finally, although sharing model parameters and
584 gradients is more privacy-protecting than sharing de-identified individual-level data (as in a centralized
585 approach), it can still pose privacy risks due to the amount of information that extremely flexible ML
586 algorithms, like deep learning models, can “memorize” about the training data (“information leakage”)

587 [96]. Additional measures such as differential privacy or learning from encrypted data can be used to
588 further reduce the risk of information leakage, but they can also increase communication costs, training
589 time, and reduce model performance [96,111]. Such countermeasures may be neither desirable nor
590 needed in the presence of proper governance and trusted collaborative relationships between data
591 partners. Despite these challenges, recent studies using federated learning with different ML
592 architectures have shown that it is possible to achieve levels of performance comparable to models
593 trained using a centralized approach [112] and better than locally trained models [113]. For example, in
594 a multi-site study using structured EHR data (e.g., laboratory data and vital signs) and chest X-ray images
595 to predict the future oxygen requirements of symptomatic patients with COVID-19 [113], the global
596 federated deep learning model – trained using data from 20 clinical sites around the world –
597 outperformed all local models that were trained at a single site using that site’s data (Table 2).
598 Federated learning is still very much an emerging and active area of research that will continue to
599 develop in the years to come [96] – not only for predicting health outcomes, but also for estimating
600 causal effects, which to our knowledge has not yet been extensively explored for doubly-robust causal
601 estimation frameworks like targeted maximum likelihood estimation [114] that are becoming
602 increasingly popular in pharmacoepidemiologic research.

603 5 Conclusions

604 Many opportunities exist for DDNs to use ML in pharmacoepidemiologic and pharmacovigilance studies.
605 From phenotyping activities to signal detection, causal inference, and forecasting activities, the use of
606 data-adaptive ML methods offers the potential to more fully capitalize on the larger amounts of
607 electronic health data made accessible through the formation of DDNs.
608 However, the siloed storage and diverse nature of databases in DDNs also create unique challenges and
609 considerations when applying ML. Many of these challenges stem from a DDN’s practical data-related

610 characteristics in terms of the modality of source data used, the extent to which sites within a DDN
611 standardize their source data, and the level of granularity with which sites in a DDN share their data. In
612 this paper, we presented several scenarios of DDNs with different characteristics (i.e., locations along
613 the spectrum) of these data-related factors, and in each case, discussed challenges that DDNs may face
614 when applying ML. We also discussed possible approaches for addressing these challenges, including
615 issues to consider and efforts that real-world DDNs have taken or are currently taking to help mitigate
616 these challenges.

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

634 The use of ML in DDNs also creates greater opportunities or impetus to address certain issues that often
635 plague the use of ML in single-database settings. For example, while ML models developed in single-site
636 settings are rarely and slowly (e.g., over years) externally validated, the collaboration of multiple data
637 partners using a CDM and standardized programs and code in DDNs greatly facilitates the ability to
638 easily and quickly (e.g., in months) externally validate ML models and identify models with greater
639 generalizability [107]. ML models developed in single-site settings also often suffer from a lack of
640 transparency in the model development process; in DDNs, however, issues of transparency and
641 reproducibility must be directly addressed due to the need for multiple data partners to work together
642 and process a common set of ML models. Finally, as researchers may observe unusual or heterogeneous
643 outputs from ML models across data partners in a DDN, the impetus to interpret ML models and explain
644 their outputs may be encountered more frequently in DDNs than in single-site settings, as data partners
645 work to recognize and rectify potential (unwanted) sources of heterogeneity, such as data errors and
646 idiosyncrasies in coding and documentation practices across sites – a task that undoubtedly becomes
647 more challenging with the use of more complex ML approaches. Taken together, by facilitating the
648 development of more robust, generalizable, reproducible, and interpretable ML models, the
649 opportunities for ML in DDNs may increase the likelihood that the resulting models – and their outputs –
650 are used to effectively enable or enhance decision making by clinicians, health care institutions, and
651 regulatory bodies.

652 In conclusion, there is great potential – and great desire – to use ML in DDNs to enhance their activities
653 in pharmacoepidemiology and pharmacovigilance. Indeed, the time is ripe for this emerging area of
654 interest among DDNs due to not only the recent methodologic advancements in the field, but also the
655 valuable groundwork that many DDNs have already laid through investments in harmonizing datasets to
656 a CDM, developing standardized processes, tools, and analytics, and building collaborative relationships
657 between data partners and with various stakeholders. The future holds much promise for the use of ML

658 in DDNs, and we expect that the utility of these data-adaptive methods for enhancing
659 pharmacoepidemiologic and pharmacovigilance studies will likely continue to increase in the years to
660 come.

661 [Statements and Declarations](#)

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

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

678 **Availability of data and material:** Not applicable

679 **Code availability:** Not applicable

680 **Ethics approval:** Not applicable

681 **Consent to participate:** Not applicable

682 **Consent for publication:** Not applicable

683 **Author contributions:** JW and ST conceptualized and drafted the article. DPA, PRR, RJD, JMR, and ST
684 critically reviewed the manuscript and provided important interpretation and critique of intellectual
685 content. All authors have given final approval of the version to be published and agree to be
686 accountable for all aspects of the work.

687

688

689 **References**

- 690 1. Evans RS. Electronic Health Records: Then, Now, and in the Future. *Yearb Med Inform.* 2016;S48–61.
- 691 2. Murray MD. Use of Data from Electronic Health Records for Pharmacoepidemiology. *Curr Epidemiol*
692 *Rep.* 2014;1:186–93.
- 693 3. Toh S, Pratt N, Klungel O, Gagne JJ, Platt RW. Distributed Networks of Databases Analyzed Using
694 Common Protocols and/or Common Data Models. *Pharmacoepidemiology* [Internet]. John Wiley & Sons,
695 Ltd; 2019 [cited 2021 Aug 1]. p. 617–38. Available from:
696 <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119413431.ch25>
- 697 4. Burn E, You SC, Sena A, Kostka K, Abedtash H, Abrahao MTF, et al. Deep phenotyping of 34,128
698 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America,
699 Europe and Asia: an international network study. *MedRxiv Prepr Serv Health Sci.*
700 2020;2020.04.22.20074336.
- 701 5. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical
702 and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and
703 quality of care. *Med Care.* 2010;48:S45-51.
- 704 6. US Food and Drug Administration. Sentinel System Five-Year Strategy, 2019-2023 [Internet]. 2019 Jan.
705 Available from: <https://www.fda.gov/media/120333/download>
- 706 7. Platt RW, Henry DA, Suissa S. The Canadian Network for Observational Drug Effect Studies (CNODES):
707 Reflections on the first eight years, and a look to the future. *Pharmacoepidemiol Drug Saf.* 2020;29
708 Suppl 1:103–7.
- 709 8. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy
710 of deep learning in medical imaging: a systematic review and meta-analysis. *Npj Digit Med.* 2021;4:1–23.
- 711 9. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning
712 Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform.* 2018;22:1589–
713 604.
- 714 10. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of
715 Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform.* 2019;7:e12239.
- 716 11. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-
717 scale structured electronic health records for disease prediction. *Npj Digit Med.* 2021;4:1–13.
- 718 12. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, et al. Clinical Concept Embeddings
719 Learned from Massive Sources of Multimodal Medical Data. *Pac Symp Biocomput Pac Symp Biocomput.*
720 2020;25:295–306.
- 721 13. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized
722 analytics pipeline for reliable and rapid development and validation of prediction models using
723 observational health data. *Comput Methods Programs Biomed.* 2021;106394.

- 724 14. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A Review of Automatic
725 Phenotyping Approaches using Electronic Health Records. *Electronics*. Multidisciplinary Digital Publishing
726 Institute; 2019;8:1235.
- 727 15. Karim ME, Pang M, Platt RW. Can We Train Machine Learning Methods to Outperform the High-
728 dimensional Propensity Score Algorithm? *Epidemiol Camb Mass*. 2018;29:191–8.
- 729 16. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-
730 world and synthetic data experiments. *Int J Epidemiol*. 2018;47:2005–14.
- 731 17. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018;319:1317–8.
- 732 18. AsPEN collaborators, Andersen M, Bergman U, Choi N-K, Gerhard T, Huang C, et al. The Asian
733 Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for
734 pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf*. 2013;22:700–4.
- 735 19. Suissa S, Henry D, Caetano P, Dormuth CR, Ernst P, Hemmelgarn B, et al. CNODES: the Canadian
736 Network for Observational Drug Effect Studies. *Open Med*. 2012;6:e134–40.
- 737 20. European Health Data Evidence Network [Internet]. ehden.eu. [cited 2021 Sep 7]. Available from:
738 <https://www.ehden.eu/>
- 739 21. Steiner JF, Paolino AR, Thompson EE, Larson EB. Sustaining Research Networks: the Twenty-Year
740 Experience of the HMO Research Network. *EGEMS Wash DC*. 2014;2:1067.
- 741 22. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current
742 state, accomplishments, and future directions. *J Clin Epidemiol*. 2021;129:60–7.
- 743 23. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data
744 Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol*
745 *Inform*. 2015;216:574–8.
- 746 24. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA’s sentinel initiative--A comprehensive approach to
747 medical product surveillance. *Clin Pharmacol Ther*. 2016;99:265–8.
- 748 25. McNeil MM, Gee J, Weintraub ES, Belongia EA, Lee GM, Glanz JM, et al. The Vaccine Safety Datalink:
749 successes and challenges monitoring vaccine safety. *Vaccine*. 2014;32:5390–8.
- 750 26. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks:
751 demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med*.
752 2016;71:57–61.
- 753 27. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping:
754 From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci*. 2018;1:53–68.
- 755 28. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from
756 electronic health record data. *Curr Epidemiol Rep*. 2018;5:331–42.

- 757 29. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for
758 patient phenotyping from electronic health records. *Sci Rep.* 2020;10:1432.
- 759 30. Ong M-S, Klann JG, Lin KJ, Maron BA, Murphy SN, Natter MD, et al. Claims-Based Algorithms for
760 Identifying Patients With Pulmonary Hypertension: A Comparison of Decision Rules and Machine-
761 Learning Approaches. *J Am Heart Assoc.* 2020;9:e016648.
- 762 31. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic
763 medical records to improve case detection: a systematic review. *J Am Med Inform Assoc JAMIA.*
764 2016;23:1007–15.
- 765 32. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthkrishnan AN, et al. Development of
766 phenotype algorithms using electronic medical records and incorporating natural language processing.
767 *BMJ. British Medical Journal Publishing Group;* 2015;350:h1885.
- 768 33. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype
769 algorithm evaluator. *J Biomed Inform.* 2019;97:103258.
- 770 34. Bate A, Hornbuckle K, Juhaeri J, Motsko SP, Reynolds RF. Hypothesis-free signal detection in
771 healthcare databases: finding its value for pharmacovigilance. *Ther Adv Drug Saf.* SAGE Publications;
772 2019;10:2042098619864744.
- 773 35. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in
774 longitudinal observational databases. *Stat Methods Med Res.* 2013;22:39–56.
- 775 36. Arnaud M, Bégau B, Thurin N, Moore N, Pariente A, Salvo F. Methods for safety signal detection in
776 healthcare databases: a literature review. *Expert Opin Drug Saf.* 2017;16:721–32.
- 777 37. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network
778 method for adverse drug reaction signal generation. *Eur J Clin Pharmacol.* 1998;54:315–21.
- 779 38. Wang SV, Maro JC, Baro E, Izem R, Dashevsky I, Rogers JR, et al. Data Mining for Adverse Drug Events
780 With a Propensity Score-matched Tree-based Scan Statistic. *Epidemiol Camb Mass.* 2018;29:895–903.
- 781 39. Wang SV, Maro JC, Gagne JJ, Patorno E, Kattinakere S, Stojanovic D, et al. A General Propensity Score
782 for Signal Identification Using Tree-Based Scan Statistics. *Am J Epidemiol.* 2021;190:1424–33.
- 783 40. Reps JM, Garibaldi JM, Aickelin U, Gibson JE, Hubbard RB. A supervised adverse drug reaction
784 signalling framework imitating Bradford Hill’s causality considerations. *J Biomed Inform.* 2015;56:356–
785 68.
- 786 41. Liu F, Jagannatha A, Yu H. Towards Drug Safety Surveillance and Pharmacovigilance: Current Progress
787 in Detecting Medication and Adverse Drug Events from Electronic Health Records. *Drug Saf.* 2019;42:95–
788 7.
- 789 42. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events
790 and medication extraction in electronic health records. *J Am Med Inform Assoc JAMIA.* 2020;27:3–12.

- 791 43. Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. unstructured: factors affecting adverse
792 drug reaction documentation in an EMR repository. *AMIA Annu Symp Proc AMIA Symp.*
793 2011;2011:1270–9.
- 794 44. Hohl CM, Kuramoto L, Yu E, Rogula B, Stausberg J, Sobolev B. Evaluating adverse drug event
795 reporting in administrative data from emergency departments: a validation study. *BMC Health Serv Res.*
796 2013;13:473.
- 797 45. Kennedy EH. Semiparametric Theory and Empirical Processes in Causal Inference. In: He H, Wu P,
798 Chen D-G (Din), editors. *Stat Causal Inferences Their Appl Public Health Res* [Internet]. Cham: Springer
799 International Publishing; 2016 [cited 2021 Sep 1]. p. 141–67. Available from:
800 https://doi.org/10.1007/978-3-319-41259-7_8
- 801 46. Schneeweiss S, Suissa S. Advanced Approaches to Controlling Confounding in
802 Pharmacoepidemiologic Studies. *Pharmacoepidemiology* [Internet]. John Wiley & Sons, Ltd; 2019 [cited
803 2021 Aug 9]. p. 1078–107. Available from:
804 <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119413431.ch43>
- 805 47. Zivich PN, Breskin A. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators.
806 *Epidemiology.* 2021;32:393–401.
- 807 48. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for
808 propensity score models. *Am J Epidemiol.* 2006;163:1149–56.
- 809 49. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification
810 methods as alternatives to logistic regression. *J Clin Epidemiol.* 2010;63:826–33.
- 811 50. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6:Article25.
- 812 51. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity
813 score adjustment in studies of treatment effects using health care claims data. *Epidemiol Camb Mass.*
814 2009;20:512–22.
- 815 52. Ju C, Combs M, Lendle SD, Franklin JM, Wyss R, Schneeweiss S, et al. Propensity score prediction for
816 electronic healthcare databases using super learner and high-dimensional propensity score methods. *J*
817 *Appl Stat.* Taylor & Francis; 2019;46:2216–36.
- 818 53. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM. Using Super Learner Prediction
819 Modeling to Improve High-dimensional Propensity Score Estimation. *Epidemiol Camb Mass.* 2018;29:96–
820 106.
- 821 54. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal
822 treatment effects. *Clin Epidemiol.* Dove Press; 2018;10:771–88.
- 823 55. Soyiri IN, Reidpath DD. An overview of health forecasting. *Environ Health Prev Med.* 2013;18:1–9.
- 824 56. Wright G, Lawrence MJ, Collopy F. The role and validity of judgment in forecasting. *Int J Forecast.*
825 1996;12:1–8.

- 826 57. Farinholt P, Park M, Guo Y, Bruera E, Hui D. A Comparison of the Accuracy of Clinician Prediction of
827 Survival Versus the Palliative Prognostic Index. *J Pain Symptom Manage*. 2018;55:792–7.
- 828 58. Saposnik G, Cote R, Mamdani M, Raptis S, Thorpe KE, Fang J, et al. JURaSSiC. *Neurology*.
829 2013;81:448–55.
- 830 59. Rojas JC, Lyons PG, Jiang T, Kilaru M, McCauley L, Picart J, et al. Accuracy of Clinicians’ Ability to
831 Predict the Need for Intensive Care Unit Readmission. *Ann Am Thorac Soc*. American Thoracic Society -
832 AJRCCM; 2020;17:847–53.
- 833 60. Sun JW, Franklin JM, Rough K, Desai RJ, Hernández-Díaz S, Huybrechts KF, et al. Predicting overdose
834 among individuals prescribed opioids using routinely collected healthcare utilization data. *PLoS ONE*.
835 2020;15:e0241083.
- 836 61. Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of Machine-Learning
837 Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid
838 Prescriptions. *JAMA Netw Open*. 2019;2:e190968.
- 839 62. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine Learning Approaches to
840 Predict 6-Month Mortality Among Patients With Cancer. *JAMA Netw Open*. 2019;2:e1915997.
- 841 63. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting Suicide
842 Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. *Am J*
843 *Psychiatry*. 2018;175:951–60.
- 844 64. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need
845 high-cost patients using nationwide clinical and claims data. *Npj Digit Med*. 2020;3:1–9.
- 846 65. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities
847 beyond structured data in analysis of electronic health records. *WIREs Comput Stat*. n/a:e1549.
- 848 66. Kong H-J. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res*. 2019;25:1–2.
- 849 67. Speech and Language Processing [Internet]. [cited 2021 Sep 13]. Available from:
850 <https://web.stanford.edu/~jurafsky/slp3/>
- 851 68. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: A methodology
852 review. *J Biomed Inform*. 2020;109:103526.
- 853 69. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction
854 applications: A literature review. *J Biomed Inform*. 2018;77:34–49.
- 855 70. projects:workgroups:nlp-wg [Observational Health Data Sciences and Informatics] [Internet]. [cited
856 2021 Sep 13]. Available from: [https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-](https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg#objective)
857 [wg#objective](https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg#objective)
- 858 71. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common Problems, Common Data
859 Model Solutions: Evidence Generation for Health Technology Assessment. *PharmacoEconomics*.
860 2021;39:275–85.

861 72. Informatics OHDS and. The Book of OHDSI [Internet]. [cited 2021 Sep 13]. Available from:
862 <https://ohdsi.github.io/TheBookOfOhdsi/>

863 73. Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How pharmacoepidemiology networks
864 can manage distributed analyses to improve replicability and transparency and minimize bias.
865 *Pharmacoepidemiol Drug Saf.* 2019;

866 74. Health Analytics Data-to-Evidence Suite (HADES) [Internet]. Observational Health Data Sciences and
867 Informatics; 2021 [cited 2021 Oct 15]. Available from: <https://github.com/OHDSI/Hades>

868 75. Klungel OH, Kurz X, de Groot MCH, Schlienger RG, Tcherny-Lessenot S, Grimaldi L, et al. Multi-centre,
869 multi-database studies with common protocols: lessons learnt from the IMI PROTECT project.
870 *Pharmacoepidemiol Drug Saf.* 2016;25 Suppl 1:156–65.

871 76. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in
872 comparative effectiveness research conducted within distributed research networks. *Med Care.*
873 2013;51:S4-10.

874 77. Jeon G, Sangaiah AK, Chen Y-S, Paul A. Special issue on Machine learning approaches and challenges
875 of missing data in the era of big data. *Int J Mach Learn Cybern.* 2019;10:2589–91.

876 78. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality
877 Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.
878 *EGEMS Wash DC.* 2016;4:1244.

879 79. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and
880 multisite data quality assessment in electronic health record-based clinical research. *Med Care.* 2012;50
881 Suppl:S21-29.

882 80. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in
883 distributed data networks. *Med Care.* 2013;51:S22-29.

884 81. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting of data
885 quality in distributed data networks. *EGEMS Wash DC.* 2015;3:1052.

886 82. Adimadhyam S, Barreto EF, Cocoros NM, Toh S, Brown JS, Maro JC, et al. Leveraging the Capabilities
887 of the FDA’s Sentinel System To Improve Kidney Care. *J Am Soc Nephrol. American Society of*
888 *Nephrology;* 2020;31:2506–16.

889 83. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through
890 evaluation of observational data quality. *J Am Med Inform Assoc.* 2021;28:2251–7.

891 84. Shi X, Li X, Cai T. Spherical Regression Under Mismatch Corruption With Application to Automated
892 Knowledge Translation. *J Am Stat Assoc. Taylor & Francis;* 2020;0:1–12.

893 85. Using Unsupervised Learning to Harmonize Data Across Data Systems | Sentinel Initiative [Internet].
894 [cited 2021 Sep 19]. Available from: [https://www.sentinelinitiative.org/methods-data-](https://www.sentinelinitiative.org/methods-data-tools/methods/using-unsupervised-learning-harmonize-data-across-data-systems)
895 [tools/methods/using-unsupervised-learning-harmonize-data-across-data-systems](https://www.sentinelinitiative.org/methods-data-tools/methods/using-unsupervised-learning-harmonize-data-across-data-systems)

896 86. Schuemie MJ, Madigan D, Ryan PB, Reich C, Suchard MA, Berlin JA, et al. Comment on “How
897 pharmacoepidemiology networks can manage distributed analyses to improve replicability and
898 transparency and minimize bias.” *Pharmacoepidemiol Drug Saf.* 2019;28:1032–3.

899 87. Arterburn D, Wellman R, Emiliano A, Smith SR, Odegaard AO, Murali S, et al. Comparative
900 Effectiveness and Safety of Bariatric Procedures for Weight Loss: A PCORnet Cohort Study. *Ann Intern
901 Med.* 2018;169:741–50.

902 88. Hurst JH, Liu Y, Maxson PJ, Permar SR, Boulware LE, Goldstein BA. Development of an electronic
903 health records datamart to support clinical and population health research. *J Clin Transl Sci.* 2020;5:e13.

904 89. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO Research Network Virtual
905 Data Warehouse: A Public Data Model to Support Collaboration. *EGEMS.* 2014;2:1049.

906 90. PCORnet® COVID-19 Common Data Model Launched, Enabling Rapid Capture of Insights on Patients
907 Infected with the Novel Coronavirus [Internet]. *Natl. Patient-Centered Clin. Res. Netw.* 2020 [cited 2021
908 Sep 20]. Available from: [https://pcor.net.org/news/pcor-net-covid-19-common-data-model-launched-
909 enabling-rapid-capture-of-insights/](https://pcor.net.org/news/pcor-net-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/)

910 91. Cocoros NM, Fuller CC, Adimadhyam S, Ball R, Brown JS, Pan GJD, et al. A COVID-19-ready public
911 health surveillance system: The Food and Drug Administration’s Sentinel System. *Pharmacoepidemiol
912 Drug Saf.* 2021;30:827–37.

913 92. Validation of Anaphylaxis Using Machine Learning | Sentinel Initiative [Internet]. [cited 2021 Sep 29].
914 Available from: [https://www.sentinelinitiative.org/methods-data-tools/methods/validation-anaphylaxis-
915 using-machine-learning](https://www.sentinelinitiative.org/methods-data-tools/methods/validation-anaphylaxis-using-machine-learning)

916 93. Validation of Acute Pancreatitis Using Machine Learning and Multi-Site Adaptation for Anaphylaxis |
917 Sentinel Initiative [Internet]. [cited 2021 Sep 21]. Available from:
918 [https://www.sentinelinitiative.org/methods-data-tools/methods/validation-acute-pancreatitis-using-
919 machine-learning-and-multi-site](https://www.sentinelinitiative.org/methods-data-tools/methods/validation-acute-pancreatitis-using-machine-learning-and-multi-site)

920 94. Carrell DS. Improving methods of identifying anaphylaxis for medical product safety surveillance
921 using natural language processing and machine learning [Internet]. 2021. Available from:
922 [https://sentinelinitiative.org/sites/default/files/documents/ICPE%20Presentation%20-
923 %20Improving%20Methods%20of%20Identifying%20Anaphylaxis%20for%20Medical%20Product%20Saf
924 ety.pdf](https://sentinelinitiative.org/sites/default/files/documents/ICPE%20Presentation%20-%20Improving%20Methods%20of%20Identifying%20Anaphylaxis%20for%20Medical%20Product%20Safety.pdf)

925 95. Extending Machine Learning Methods Development in Sentinel: Follow-up Analyses for Anaphylaxis
926 Algorithm and Formalization of a General Phenotyping Framework (Phase 3) | Sentinel Initiative
927 [Internet]. [cited 2021 Sep 29]. Available from: [https://www.sentinelinitiative.org/methods-data-
928 tools/methods/extending-machine-learning-methods-development-sentinel-follow-analyses](https://www.sentinelinitiative.org/methods-data-tools/methods/extending-machine-learning-methods-development-sentinel-follow-analyses)

929 96. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with
930 federated learning. *Npj Digit Med.* 2020;3:1–7.

931 97. NOTE NLP table · Issue #85 · OHDSI/CommonDataModel [Internet]. GitHub. [cited 2021 Sep 4].
932 Available from: <https://github.com/OHDSI/CommonDataModel/issues/85>

933 98. NLPTools/Wrappers at master · OHDSI/NLPTools [Internet]. GitHub. [cited 2021 Sep 21]. Available
934 from: <https://github.com/OHDSI/NLPTools>

935 99. Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend Medical: a Named Entity Recognition and
936 Relationship Extraction Web Service. ArXiv191007419 Cs [Internet]. 2019 [cited 2021 Sep 6]; Available
937 from: <http://arxiv.org/abs/1910.07419>

938 100. Map clinical notes to the OMOP Common Data Model and healthcare ontologies using Amazon
939 Comprehend Medical [Internet]. Amaz. Web Serv. 2019 [cited 2021 Sep 21]. Available from:
940 [https://aws.amazon.com/blogs/machine-learning/map-clinical-notes-to-the-omop-common-data-
941 model-and-healthcare-ontologies-using-amazon-comprehend-medical/](https://aws.amazon.com/blogs/machine-learning/map-clinical-notes-to-the-omop-common-data-model-and-healthcare-ontologies-using-amazon-comprehend-medical/)

942 101. Representation of Unstructured Data Across Common Data Models | Sentinel Initiative [Internet].
943 [cited 2021 Sep 21]. Available from: [https://www.sentinelinitiative.org/methods-data-
944 tools/methods/representation-unstructured-data-across-common-data-models](https://www.sentinelinitiative.org/methods-data-tools/methods/representation-unstructured-data-across-common-data-models)

945 102. Advancing Scalable Natural Language Processing Approaches for Unstructured Electronic Health
946 Record Data | Sentinel Initiative [Internet]. [cited 2021 Sep 21]. Available from:
947 [https://www.sentinelinitiative.org/methods-data-tools/methods/advancing-scalable-natural-language-
948 processing-approaches-unstructured](https://www.sentinelinitiative.org/methods-data-tools/methods/advancing-scalable-natural-language-processing-approaches-unstructured)

949 103. Improving Methods for Identifying Social, Behavioral, and Clinical Factors in Doctors' Notes in
950 Electronic Health Records [Internet]. 2019 [cited 2021 Sep 21]. Available from:
951 [https://www.pcori.org/research-results/2019/improving-methods-identifying-social-behavioral-and-
952 clinical-factors-doctors%E2%80%99](https://www.pcori.org/research-results/2019/improving-methods-identifying-social-behavioral-and-clinical-factors-doctors%E2%80%99)

953 104. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a
954 standardized framework to generate and evaluate patient-level prediction models using observational
955 healthcare data. *J Am Med Inform Assoc JAMIA*. 2018;25:969–75.

956 105. Wang Q, Reps JM, Kostka KF, Ryan PB, Zou Y, Voss EA, et al. Development and validation of a
957 prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale
958 in the OHDSI network. *PLOS ONE*. Public Library of Science; 2020;15:e0226718.

959 106. Kim C, You SC, Reps JM, Cheong JY, Park RW. Machine-learning model to predict the cause of death
960 using a stacking ensemble method for observational data. *J Am Med Inform Assoc JAMIA*.
961 2021;28:1098–107.

962 107. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a
963 large-scale external validation approach for patient-level prediction in an international data network:
964 validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. *BMC
965 Med Res Methodol*. 2020;20:102.

966 108. Williams RD, Reps JM, Kors JA, Ryan PB, Steyerberg E, Verhamme KM, et al. Using iterative pairwise
967 external validation to contextualize prediction model performance: A use case predicting 1-year heart-
968 failure risk in diabetes patients across five data sources. *Drug Saf*. 2022;

- 969 109. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated Learning with Non-IID Data.
970 ArXiv180600582 Cs Stat [Internet]. 2018 [cited 2021 Sep 28]; Available from:
971 <http://arxiv.org/abs/1806.00582>
- 972 110. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated Optimization in Heterogeneous
973 Networks. ArXiv181206127 Cs Stat [Internet]. 2020 [cited 2021 Sep 27]; Available from:
974 <http://arxiv.org/abs/1812.06127>
- 975 111. Li T, Sahu AK, Talwalkar A, Smith V. Federated Learning: Challenges, Methods, and Future
976 Directions. IEEE Signal Process Mag. 2020;37:50–60.
- 977 112. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research
978 with federated learning. Npj Digit Med. 2021;4:1–8.
- 979 113. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting
980 clinical outcomes in patients with COVID-19. Nat Med. 2021;1–9.
- 981 114. Laan MJ van der, Rubin D. Targeted Maximum Likelihood Learning. Int J Biostat [Internet]. De
982 Gruyter; 2006 [cited 2021 Sep 30];2. Available from:
983 <https://www.degruyter.com/document/doi/10.2202/1557-4679.1043/html>
- 984

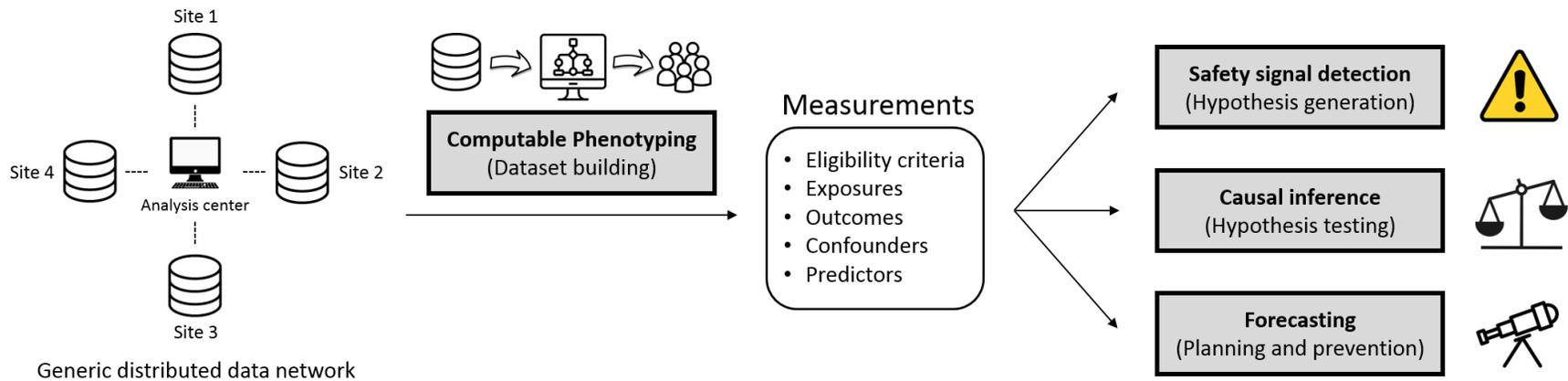
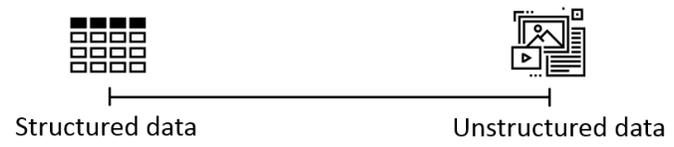


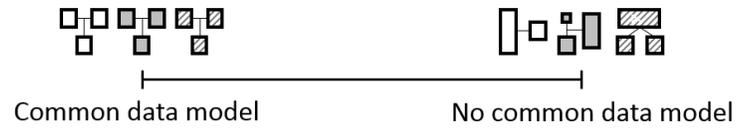
Figure 1. Key domains of activities performed by distributed data networks conducting studies in pharmacoepidemiology and pharmacovigilance

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 grey rectangles represent the key domains of activities performed by distributed data networks that conduct studies in pharmacoepidemiology and pharmacovigilance.

Modality of source data



Degree of data standardization



Granularity of shared data

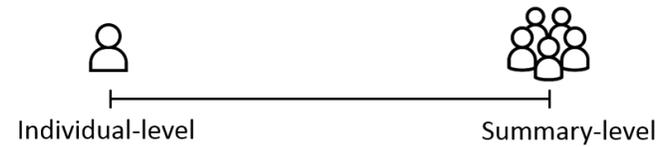


Figure 2. Practical data-related factors of distributed data networks

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.

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: Roux-en-Y gastric bypass (RYGB), sleeve gastrectomy (SG), and adjustable gastric banding (AGB) [87]	41 US health systems in the National Patient-Centered Clinical Research Network (n=46,510)	<ul style="list-style-type: none"> • LASSO used to simultaneously select features and estimate parameters for propensity score models • As 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 	<ul style="list-style-type: none"> • RYGB associated with greater weight loss than SG or AGB at 1-, 3-, 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 versus 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]	Kaiser Permanente Washington (KPWA, n=239) and Kaiser Permanente Northwest (KPNW, n=277)	<ul style="list-style-type: none"> • 5 machine learning algorithms (logistic regression, elastic net, Bayesian Additive Regression Trees [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 	<ul style="list-style-type: none"> • Adding features derived from unstructured clinical text improved the performance of phenotyping algorithms compared to 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 (HT) within 30 days of an initial acute ischemic stroke [105]	11 databases from three countries (USA, Germany, and Japan) within the OHDSI network (n=6,848,096)	<ul style="list-style-type: none"> • LASSO used to simultaneously select features and estimate parameters to predict risk of HT 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 	<ul style="list-style-type: none"> • 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 ranging from 0.60-0.78
To develop and validate a machine	2 databases from South Korea with	<ul style="list-style-type: none"> • Two-level stacking ensemble used to predict cause of death within 60 days, which consisted of a meta-learner 	<ul style="list-style-type: none"> • In the development database (n=174,747), internal validation of the stacking ensemble had an AUC of 0.95

<p>learning model to predict cause of death (within 60 days) from a patient's last medical check-up [106]</p>	<p>cause-of-death data (n=903,812), 3 US databases with no cause-of-death data (n=2,397,494)</p>	<p>that used the outputs from a collection of base learners as inputs to make the final prediction</p> <ul style="list-style-type: none"> • 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 	<p>in the hold-out test set; external validation of the stacking ensemble had an AUC of 0.89</p> <ul style="list-style-type: none"> • 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 US; these discrepancies suspected to be at least partly attributable to differences between the countries in the definition of heart disease death
<p>To develop and validate a prognostic model predicting 1-year risk of incident heart failure (HF) in patients with type 2 diabetes initiating a second pharmacotherapy for type 2 diabetes [108]</p>	<p>5 US databases (n=403,187)</p>	<ul style="list-style-type: none"> • LASSO used to simultaneously select features and estimate parameters to predict 1-year risk of incident HF • 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 	<ul style="list-style-type: none"> • Internal validation of site-specific models had AUC ranging from 0.64-0.71 for baseline (age and sex) models and AUC ranging from 0.73-0.81 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
<p>To predict clinical outcomes in patients with COVID-19 who present to the emergency department [113]</p>	<p>20 clinical institutes from various regions around the world (n=16,148)</p>	<ul style="list-style-type: none"> • 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 (EMR) data and 1 chest x-ray image, where EMR 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 	<ul style="list-style-type: none"> • For predicting 24-hour oxygen treatment, global model outperformed all local models: average AUC for the global versus locally trained models was 0.92 versus 0.80 (16% improvement), and AUC for global model also provided average increase in generalizability of 38% compared to 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

Abbreviations: AUC = area under the receiver operating curve (possible values from 0 to 1, values closer to 1 indicate better performance); EHR = electronic health record; LASSO = Least Absolute Shrinkage and Selection Operator; NLP = natural language processing; OHDSI = Observational Health Data Sciences and Informatics; OR = odds ratio