

INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL
REQUIREMENTS FOR PHARMACEUTICALS FOR HUMAN USE

ICH HARMONISED GUIDELINE

**General Principles on Plan, Design and Analysis of
Pharmacoepidemiological Studies That Utilize Real-World
Data for Safety Assessment of Medicines**

M14

Draft version
Endorsed on 21 May 2024
Currently under public consultation

At Step 2 of the ICH Process, a consensus draft text or guideline, agreed by the appropriate ICH Expert Working Group, is transmitted by the ICH Assembly to the regulatory authorities of the ICH regions for internal and external consultation, according to national or regional procedures.

M14
Document History

Code	History	Date
M14	Endorsement by the Members of the ICH Assembly under <i>Step 2</i> and release for public consultation.	21 May 2024

Legal notice: This document is protected by copyright and may, with the exception of the ICH logo, be used, reproduced, incorporated into other works, adapted, modified, translated, or distributed under a public license provided that ICH's copyright in the document is acknowledged at all times. In case of any adaption, modification or translation of the document, reasonable steps must be taken to clearly label, demarcate or otherwise identify that changes were made to or based on the original document. Any impression that the adaption, modification, or translation of the original document is endorsed or sponsored by the ICH must be avoided.

The document is provided "as is" without warranty of any kind. In no event shall the ICH or the authors of the original document be liable for any claim, damages or other liability arising from the use of the document.

The above-mentioned permissions do not apply to content supplied by third parties. Therefore, for documents where the copyright vests in a third party, permission for reproduction must be obtained from this copyright holder.

General Principles on Plan, Design and Analysis of Pharmacoepidemiological Studies That Utilize Real-World Data for Safety Assessment of Medicines

M14

ICH Consensus Guideline

Table of Contents

1	INTRODUCTION.....	1
1.1	Objectives.....	1
1.2	Background.....	1
1.3	Scope.....	2
1.4	Studies Conducted for Purposes other than the Safety Assessment of Medicines	3
2	GENERAL PRINCIPLES.....	3
3	FRAMEWORK FOR GENERATING ADEQUATE EVIDENCE USING REAL- WORLD DATA	4
4	INITIAL DESIGN AND FEASIBILITY.....	5
4.1	Research Question.....	5
4.2	Feasibility Assessment(s)	6
5	PROTOCOL DEVELOPMENT.....	8
5.1	Study Design	9
5.2	Data Sources	10
5.2.1	<i>Appropriateness of Data Sources in Addressing Safety Questions of Interest</i>	<i>11</i>
5.2.2	<i>Characteristics of Major Data Sources</i>	<i>12</i>
5.2.3	<i>Data Standardization.....</i>	<i>16</i>
5.2.4	<i>Missing Data.....</i>	<i>17</i>
5.2.5	<i>Data Quality</i>	<i>17</i>
5.2.6	<i>Data Collection and Data Source Sections in the Study Protocol.....</i>	<i>17</i>
5.3	Target/Study Population.....	18
5.4	Exposures, Outcomes, Covariates	18
5.4.1	<i>Exposure.....</i>	<i>20</i>
5.4.2	<i>Outcome.....</i>	<i>22</i>
5.4.3	<i>Covariates</i>	<i>23</i>
5.5	Bias and Confounding	25
5.5.1	<i>Selection Bias</i>	<i>25</i>
5.5.2	<i>Information Bias</i>	<i>26</i>
5.5.3	<i>Immortal Time Bias.....</i>	<i>26</i>
5.5.4	<i>Confounding.....</i>	<i>26</i>
5.6	Validation	27
6	DATA MANAGEMENT	28
6.1	Data Holder	28

6.2	Researchers.....	29
7	ANALYSIS.....	29
7.1	Statistical Analysis	30
	7.1.1 Primary Analyses.....	30
	7.1.2 Missing Data.....	31
	7.1.3 Sensitivity Analyses	31
8	REPORTING AND SUBMISSION.....	32
8.1	Reporting of Adverse Events, Adverse Drug Reactions, and Product Quality Complaints	32
8.2	Formatting and Content of Study Documents for Submission to Regulatory Authorities	32
9	DISSEMINATION AND COMMUNICATION OF STUDY MATERIALS AND FINDINGS.....	32
10	STUDY DOCUMENTATION AND RECORD RETENTION	33
11	CONSIDERATIONS IN SPECIFIC POPULATIONS.....	34
11.1	Pregnancy Studies.....	34
12	GLOSSARY.....	35
13	REGULATORY GUIDELINES REFERENCED	42
14	NON-REGULATORY GUIDELINES REFERENCED	43
15	REFERENCES.....	45

1 **1 Introduction**

2 **1.1 Objectives**

3 The purpose of this document is to recommend international standards for, and promote
4 harmonization of, the general principles on planning, designing, and analyzing observational
5 (non-interventional) pharmacoepidemiological studies that utilize fit-for-purpose data for
6 safety assessment of medicines (drugs, vaccines, and other biological products).

7 This document outlines recommendations and high-level best practices for the conduct of these
8 studies, to streamline the development and regulatory assessment of study protocols and
9 reports. These recommendations and practices also seek to improve the ability of the study
10 protocol and/or results to be accepted across health authorities and support decision-making in
11 response to study results. The Glossary defines several terms for the purpose of this guideline.
12 Terms that appear in *bold italic* type upon first use are defined in the Glossary.

13 **1.2 Background**

14 Pharmacoepidemiological studies have long been a source of data and evidence to support the
15 evaluation of the post-marketing safety of approved *medicines*.

16 Signals can arise from a wide variety of data sources. This potentially includes all clinical and
17 scientific information concerning the use of medicines and the outcome of this use, such as
18 product quality, non-clinical and clinical data (including pharmacovigilance and
19 pharmacoepidemiological data). Epidemiological studies are a key component in the detection,
20 characterization and evaluation of safety concerns or signals and may be descriptive or
21 inferential.

22 Generation of robust evidence to be used for regulatory purposes relies on the reliability and
23 relevance of the data and the application of sound pharmacoepidemiological methods to
24 analyze such data. The use of pharmacoepidemiological studies for regulatory decision-making
25 has increased globally, and multiple guidelines and best practice documents have been
26 developed by health authorities and professional societies. Many countries and regions have
27 published guidelines related to general principles of planning and designing such studies
28 mainly for the purpose of safety assessment of a medicine. In addition, frameworks for study
29 design and conduct are being developed by non-governmental groups, such as The Sentinel
30 Innovation Center with the PRINCIPLED framework and ISPE/ISPOR's HARmonized

31 Protocol Template to Enhance Reproducibility (HARPER) Initiative, which provide additional
32 detail that is beyond the scope of this guideline [1, 5].

33 **1.3 Scope**

34 While recognizing that there may be slight differences between regions with regard to what
35 constitutes *real-world data* (RWD), this guideline includes recommendations for studies
36 utilizing RWD conducted for the purposes of evaluating post-marketing safety of medicinal
37 products. At times, RWD sources alone may be insufficient to answer the research question of
38 interest and researchers will gather additional data for the purposes of the study. For the purpose
39 of this guidance, we refer to data collected for the specific study as primary data collection.
40 Because primary data collection may be relevant to observational studies using RWD, when
41 relevant, this guideline also includes considerations for primary data collection.

42 It is beyond the scope of this document to provide guidance on whether a clinical trial or a
43 pharmacoepidemiological study is the most appropriate approach, nor is it intended as a
44 comprehensive source of knowledge for pharmacoepidemiological methods. Rather, the intent
45 is to harmonize regulatory guidance documents for the design, planning and execution of
46 pharmacoepidemiological studies, and to facilitate regulatory review. Parties can also consider,
47 as relevant, best practice guidances from other sources to the extent not covered in regulatory
48 guidance (see Non-regulatory Guidelines Referenced).

49 The following study types are out of scope:

- 50 • Pharmacovigilance studies using spontaneous reports obtained from national or global
51 databases (e.g., pharmacovigilance systems at national level);
- 52 • Studies involving treatment assignment, including randomized controlled trials, pragmatic
53 trials, single arm clinical trials with treatment assignment defined per protocol, and trials
54 using external comparators; and
- 55 • Studies collecting and analyzing *patient experience data*.

56 Collecting patient experience data may be a valuable component for post-marketing safety
57 studies to inform on aspects such as notable events, perspectives, needs, and priorities. While
58 a detailed guidance on this is beyond the scope of this guideline, several regulatory guidances
59 have been developed (see Section 13, Regulatory Guidelines Referenced). When studies
60 include patient experience data, the researcher may consult relevant published recommendation
61 for additional information.

62 Considering the evolving nature of pharmacogenomics, artificial intelligence (AI), and other
63 emerging technologies relevant to the use of RWD, this guideline does not address those topics.

64 **1.4 Studies Conducted for Purposes other than the Safety Assessment of Medicines**

65 The principles presented in this document provide recommendations that may be applicable to
66 post-market pharmacoepidemiological studies conducted for purposes other than evaluation of
67 the safety of medicines, such as utilization and effectiveness studies. The basic principles
68 presented in this guideline may be relevant to these studies when real-world data elements are
69 included.

70 **2 General Principles**

71 The safety profile of a medicine reflects an evolving body of knowledge extending from
72 preclinical investigations through the post-approval lifecycle. Post-approval
73 pharmacoepidemiological safety studies complement other sources of information to provide
74 a better picture of the benefit-risk profile of a medicine as used in clinical practice.

75 The guideline describes a stepwise process, although the various steps of study design and data
76 source selection are iterative. The process starts with articulating the research question;
77 conducting a systematic process to identify the study population, *exposure*, outcome, and
78 covariates; identifying minimal data requirements to guide feasibility assessment; assessing the
79 representativeness of the data source to the target population; and considering sources of
80 potential *bias* and *confounding*. After an appropriate data source and/or data collection
81 approach has been identified, the process involves further refining the design, which includes
82 approaches to address study validity. The fit-for-purpose evaluation section of the guideline
83 describes the integration of these activities. Throughout the process, the underlying rationale
84 and justification for exposure, outcome, and confounder definitions, analysis, data
85 management, study implementation, reporting, submission, dissemination of results, and other
86 key decisions should be documented. All operational aspects should be clear and transparent.

87 In this guideline we refer to “researcher” as those responsible for designing and executing the
88 study; this may be a regulatory agency, sponsor, contract research organization, academic
89 group, or others. Sponsors of marketing applications and marketing authorization holders are
90 ultimately responsible for all aspects of post-marketing safety studies submitted to regulators.

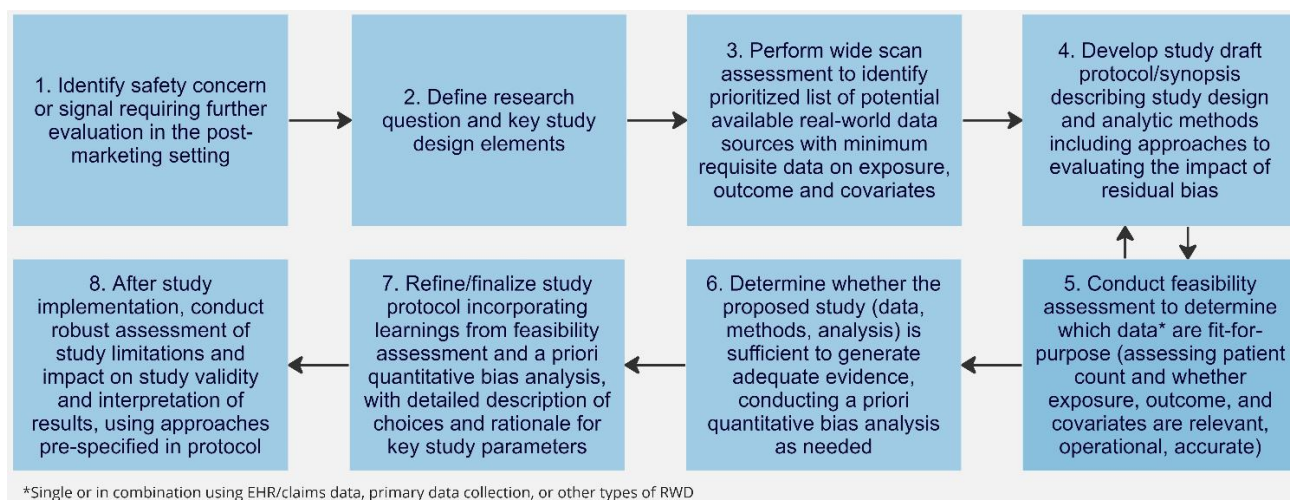
91 3 Framework for Generating Adequate Evidence using Real-World Data

92 The strength of the study generated evidence submitted in support of a regulatory decision
93 depends on the research design and methodology in addition to the relevance and reliability of
94 the underlying data. Within the framework for generating adequate evidence (Figure 1), the
95 research question should be established first, then the study design and data source(s) most
96 appropriate for addressing those questions are determined (2). Researchers should avoid
97 designing a study that conforms to a specific data source, because a specific data source may
98 restrict the options for study design and limit the inferences that can be drawn. In general, to
99 determine if the evidence that will be generated is adequate to answer the research question,
100 the framework should include an integrated assessment of (1) *data relevance* and *data*
101 *reliability*, (2) appropriateness of the study design and analytic methods, and (3) a
102 qualitative/quantitative robust assessment of study limitations and their impact on the ultimate
103 validity and reliability of the resulting evidence and the interpretation of findings. Integrated
104 assessment of whether the evidence generated through the study is adequate should be
105 considered both during protocol development with a feasibility assessment (e.g., discussion of
106 the impact of theoretical concerns, consideration of possible sources of bias and their potential
107 impact on study validity) and after study implementation with sensitivity analyses pre-specified
108 in the protocol. *Quantitative bias assessments (analyses)* may be employed either *a priori* for
109 feasibility assessment, or to facilitate interpretation of study results, or for both purposes. All
110 three components considered simultaneously can enable a decision on whether the study, if
111 executed according to the protocol, can generate adequate evidence to address the specific
112 regulatory question. Studies involving user-generated health data extracted from other sources
113 (e.g., websites, blogs, social media, chat rooms) may not be adequate, but they may provide
114 supportive data to generate hypotheses and contextualize the study results.

115 Although **Figure 1** depicts a linear process, consideration and evaluation of evidence that is
116 adequate should be iterative [2]. Researchers are encouraged to discuss the attributes of a
117 particular study with the regulatory agency early in the planning process. The ensuing sections
118 of this guideline outline the necessary elements of a study protocol that will allow for a validity
119 assessment.

120

121 **Figure 1: A framework for generating adequate evidence using fit-for-purpose real-**
 122 **world data to address regulatory questions on the safety of medicines.**



123

124 **4 Initial Design and Feasibility**

125 **4.1 Research Question**

126 The research question is a concise statement of the study purpose and the prespecified
 127 hypotheses to be tested; the purpose of the study may also be to generate hypotheses for future
 128 research. The research question may be formulated by use of the population, intervention
 129 (exposure in the case of non-interventional studies), comparator, outcome, and timing (PICOT)
 130 template. In the case of non-interventional studies, “intervention” can be considered the same
 131 as an exposure. The specific question should be formulated after a review of the literature to
 132 identify and understand any knowledge gaps, strengths and weaknesses of prior studies, the
 133 expected magnitude of effect, and important confounding factors. When defining the research
 134 question, researchers should provide a clear rationale on how it will be addressed by the study
 135 objectives. In the protocol, researchers should document and support decisions about the study
 136 design and the types of data required/available. Careful formulation of the research question
 137 will highlight unknowns that will need to be addressed through information derived from the
 138 feasibility assessment and this information may further refine the question and drive protocol
 139 development. Researchers may also consider a principled framework for study design and
 140 estimation of the risks of a medicine, such as the *target trial* approach or the *estimands*
 141 framework as they initiate work on the research question and conduct initial design and
 142 feasibility analyses [3, 5].

143 **4.2 Feasibility Assessment(s)**

144 A feasibility assessment is a systematic process to identify fit-for-purpose data to address a
 145 specific research question and to obtain information on the statistical precision of a potential
 146 study without evaluating outcomes for treatment arm. When conducting a feasibility
 147 assessment, a key goal is to describe and compare the reliability and relevance of the data
 148 sources assessed for the research question without evaluating outcomes associated with the
 149 medicine under evaluation. Additional detail on potential strengths and limitations of data
 150 sources is provided in Section 5, Protocol Development.

151 Feasibility assessments should be structured in at least two phases:

- 152 • An initial scan to determine whether data are available, likely sufficient, and to narrow
 153 down data source options; and
- 154 • A subsequent, more comprehensive feasibility assessment of the candidate data sources.

155 After the research question and design elements are established, researchers should specify the
 156 minimum criteria required to address the key design elements specific to the research question.
 157 This task will require an understanding of RWD source characteristics and the clinical context.

158 Design elements to consider include:

- 159 • Data needed to understand and define the study population, exposure, comparison groups,
 160 outcomes and covariates;
- 161 • Minimum length of follow-up to observe outcome(s);
- 162 • The targeted sample size/event rate and expected study precision;
- 163 • Geographic region(s) of interest; and
- 164 • When feasible, information about the health care system including method of diagnosis,
 165 preferred medicines, formulary coverage and prescribing practices.

166 In the early stages of designing a non-interventional study, expectations regarding access to
 167 patient level or analytic data sets should be clarified. Sponsors should obtain any required third-
 168 party agreements to access relevant patient-level or analytic data that will be required by the
 169 regulatory authority for submission.

170 Other important elements related to the feasibility assessments can include:

- 171 • Whether appropriate codes for a diagnosis are available, especially for rare diseases;
- 172 • Whether laboratory confirmation of a diagnosis and/or access to medical records are
 173 necessary to validate outcomes or exposures; and

- 174 • Whether evidence for the validity of coding algorithms exists.

175 Depending on the research question, it may be appropriate to specify other important criteria,
 176 such as the ability to collect additional information to complement records in the data sources,
 177 or link data sources to other types of data (e.g., vital records, cancer registries, vaccine
 178 registries). At this point in the initial scan step, it will be possible to identify data sources that
 179 are most likely to satisfy the criteria the researcher has specified as important to answer the
 180 research question. In some cases, it is possible for the researcher to complete this initial scan
 181 step relying on published information, data source descriptions, catalogues of metadata, and
 182 occasionally, simple descriptive counts available from the data source.

183 Once a manageable number of available data sources have been identified as potential
 184 candidates for utilization in the study, an in-depth feasibility assessment should be conducted.
 185 In some instances, fit-for-purpose data will be identified during the initial feasibility scan, in
 186 which case the detailed step will apply to the databases under consideration. In the detailed
 187 feasibility step, the researcher can verify that the specific data needed for the key design criteria
 188 are available and that there is sufficient evidence of validity and completeness of the minimal
 189 design elements in the specific data source.

190 When selecting a data source, data recency, frequency of data refresh, completeness of follow-
 191 up from exposure to outcome should be considered. In addition, the possibility to submit data
 192 files generated during conduct of the study to relevant regulatory agencies may need to be
 193 determined. Other factors in data source selection may be prior experience with the data, as
 194 there may be a trade-off between the time needed to address these factors versus the urgency
 195 of obtaining study results.

196 Analyses that evaluate the potential impact of missingness of data may be conducted to further
 197 evaluate the feasibility of conducting a study in a given data source. For example, in a study
 198 evaluating the association between hormonal contraceptives and thromboembolism, the impact
 199 of missing smoking status information may be evaluated by a review of the literature to
 200 determine the association of smoking with exposure and outcome, and then using quantitative
 201 bias approaches to evaluate the impact on the study validity for a range of desired effect
 202 estimates. A variety of information sources are used to complete this step evaluating the
 203 narrowed down list of data sources, and it may often be valuable to request specific information
 204 from the *data holder* (e.g., number of patients exposed, incidence rate of outcome to conduct
 205 sample size calculations, availability of covariates, and other queries of the data to verify the

206 data source is fit-for-purpose).

207 After the detailed evaluation is complete, the data sources are compared, and a data source(s)
208 can be selected for the study. Occasionally, at any of the steps, it will be apparent that a specific
209 data source is not suitable to address the research question. In these circumstances, the
210 researcher may conduct a feasibility assessment for primary data collection. This assessment
211 typically includes physician and site queries, including information about the patient
212 population, to determine if a sufficient number of participants can be enrolled and followed for
213 the appropriate timeframe to yield meaningful answers to the research question. Whenever
214 primary data collection studies are proposed, the researcher should consider the time to set up
215 the study which includes time to select sites, undergo ethical approval, enroll, and follow
216 participants, and produce results, and whether this timing is acceptable.

217 In addition, the specification of an appropriate comparator group (or time period) is a critical
218 part of the study design and an important consideration in the feasibility assessments. The
219 impact that policies for medical or medication coverage could have on the observed level of
220 disease severity of the exposed group and the comparator group must be considered, as should
221 the availability of concurrent comparator data. However, in some situations (such as rare
222 disease population studies) a historical or former *standard of care* comparator may be
223 considered. Regulatory guidances provide additional information on the characteristics of an
224 appropriate comparator.

225 Feasibility assessments are used as context for design decisions in the protocol. In discussion
226 with, and where required by a regulatory authority, submission of the feasibility assessment
227 report can either be a standalone document, or an annex to the protocol. This report should
228 describe the data sources evaluated when designing the study, including results from feasibility
229 evaluations or exploratory analyses of those data sources. Sponsors should provide a
230 justification for selecting or excluding relevant data sources from the study.

231 The final approach should comply with applicable regulatory requirements. Detailed
232 frameworks, templates, and checklists for conducting feasibility assessments are available in
233 scientific publications.

234 **5 Protocol Development**

235 The design and conduct of pharmacoepidemiological safety studies typically require the
236 participation of subject matter experts. An experienced, multidisciplinary study team with the

237 appropriate expertise is crucial to the successful execution of a safety study and the protocol
238 should include a description of the expertise and credentials of the study team. These personnel
239 provide essential input in a number of areas, including:

- 240 • Development of exposure, outcome and covariate definitions, with appropriate medical
241 expertise to understand disease manifestation, causal pathways, and current clinical
242 practices;
- 243 • The unique features of existing electronic healthcare data sources based on their
244 intended purpose and methods for collecting data;
- 245 • Disease area billing and coding practices;
- 246 • Specific characteristics around primary data collection; and
- 247 • Addressing potential data privacy and security concerns raised when accessing health
248 care data.

249 Completeness of data capture, bias in the assessment of exposure, outcome and covariates,
250 variability among data sources, the impact of changes over time in the data, governance and
251 conditions of access to data, and the healthcare system of the country or region covered by the
252 database are important elements that can affect the choice of the data source(s) for the study
253 and need to be addressed in the study protocol.

254 **5.1 Study Design**

255 Pharmacoepidemiological safety studies usually aim to estimate the incidence of an adverse
256 outcome in a population of interest and to evaluate its association with exposure to a medicine.

257 Several study designs are commonly used in observational pharmacoepidemiological safety
258 studies, including cohort, case-control, and self-controlled studies. The selection of the most
259 appropriate study design depends on multiple factors including the research question of interest
260 and what is known about the postulated relationship between exposure to a medicine and the
261 specific safety outcomes of interest (e.g., acute vs. latent outcome, biologic *plausibility*).

262 Identifying the appropriate comparator population (designed to represent the counterfactual
263 experience) is a critical element of study design. Examples of comparators may include users
264 of other medicines, non-users, historical controls, or the patient themselves in self-controlled
265 designs. Considerations for comparator selection may include the specific indication within a
266 disease, contraindications, disease severity or comorbidity, and the treatment sequence. It is
267 important to maximize and evaluate the comparability of the exposed and comparator

268 populations to reduce issues related to confounding by indication.

269 Researchers should discuss their rationale for selecting a particular study design in the study
270 protocol and final report. Researchers should also consider developing graphical
271 representations (such as a study design diagram) to clarify the analysis plan and time
272 components such as inclusion period, lookback period, follow-up period, overall study period.
273 Visualization of design details helps to clarify and communicate the study design to a broad
274 audience of decision makers [3]. The proposed study design should be discussed with health
275 authorities early in the process to ensure that the proposed study may provide adequate
276 evidence for regulatory decision-making.

277 After initial feasibility analyses, all essential elements of study design, analysis, conduct, and
278 reporting should be prespecified. For each study element, the protocol and final study report
279 should describe how that element was ascertained from the selected data source in studies
280 utilizing secondary data, including applicable validation studies.

281 **5.2 Data Sources**

282 Before using any data source in support of regulatory decision-making, sponsors should
283 consider whether the data are fit-for-purpose by assessing the data's relevance and reliability.
284 For the purposes of this guidance, the term relevance includes the availability of key data
285 elements (patient characteristics, exposures, outcomes) and a sufficient number of
286 representative patients for the study (target population), and the term reliability includes *data*
287 *accuracy, completeness, provenance, and traceability*. The protocol should provide discussion
288 and documentation of these key data characteristics.

289 Several data source characteristics need to be considered in pharmacoepidemiological studies,
290 as they may affect the study design and the interpretation of the results. These include
291 differences in coding systems used across databases, standardization of data elements, and
292 settings of care captured (e.g., primary, hospital, specialty, rehabilitation). Patients, providers,
293 or healthcare systems may have different motivations (monetary, social, or otherwise) for data
294 collection or participation, and billing practices for reimbursement, which may impact the
295 characteristics of the underlying data and further inform study design and interpretation.

296 In recent years, federated networks of RWD sources have been developed in various regions.
297 When utilizing multiple data sources, either as a network or through data linkage, researchers
298 should consider the steps taken to harmonize data across institutions or data sources (see
299 *Federated Data Networks*). Some of these networks have been specifically designed to support

300 scientific evaluations and regulatory decision-making, allowing a growing number of studies
301 to include data from these federated networks, often from different countries. It is essential to
302 understand the strengths and limitations of the chosen data source(s).

303 *5.2.1 Appropriateness of Data Sources in Addressing Safety Questions of Interest*

304 Researchers should demonstrate an understanding of the data source(s) and its appropriateness
305 to address specific research questions. This understanding of the chosen data source(s)
306 including the relevance and reliability of the data to address the specific research question, in
307 conjunction with an appropriate study design and analysis, is key to providing accurate
308 evidence. During development of the protocol, as informed by the feasibility assessment(s),
309 researchers should describe the following key aspects of the proposed data source(s) to support
310 the demonstration of their relevance, the selection rationale and how they might affect the
311 generalizability of the study results to the targeted patient population:

- 312 • How well the selected data source captures study elements (e.g., whether a variable is
313 captured, and if so, the degree of completeness);
- 314 • The capability to validate the outcome and other key study elements (e.g., exposures,
315 key covariates, inclusion/exclusion criteria);
- 316 • The historical experience with use of the selected data source for research purposes,
317 including references for publications citing relevant previous use for
318 pharmacoepidemiology studies which may demonstrate fit-for-use characteristics or
319 other elements to support use of the data source for the proposed study;
- 320 • Time to data availability, frequency of data refresh;
- 321 • The relevant healthcare system factors, such as medication tiering (e.g., first-line,
322 second-line), formulary decisions, and patient coverage, can influence the degree to
323 which patients on a given therapy in one health care system might differ in disease
324 severity, or other characteristics, from patients on the same therapy in another
325 healthcare system;
- 326 • The key patient characteristics which might act as potential confounders, including age,
327 socio-economic status, health conditions, risk factors for the outcome, health system
328 (e.g., private or public/governmental healthcare); and
- 329 • Potential limitations of the data source for addressing the research question.

330 **5.2.2 Characteristics of Major Data Sources**

331 Regardless of the data source(s) used, information on the context of the evidence generation
332 should be obtained (e.g., geographic location, setting in which the data were generated, period
333 during which the data were collected, and demographic information such as the age and sex
334 distribution of populations included in the data source). Examples include data derived from
335 EHRs, *administrative healthcare claims data (claims data)*, data from patient registries,
336 patient-generated data, and data gathered from other sources that can inform on health status,
337 such as interviews, mail surveys, computerized or mobile-application questionnaires,
338 measurements through digital health technologies (see [Digital Health Technologies](#)). Although
339 there are regional differences, such as medical practice, below are general considerations for
340 common data types.

341 **Electronic Health Record (EHR) Data**

342 *Electronic Health Record (EHR)* data are captured by healthcare institutions, and these data
343 reflect episodic care as captured within that specific institution and may not reflect the patient's
344 complete medical history, because they may miss data from other settings of care. Given that
345 components and formats of data may differ among medical institutions, standardization of data
346 formats is often a major issue in a study when integrating data from multiple institutions.

347 Key clinical information are often unstructured data within EHRs, either as free text fields
348 (such as healthcare practitioner notes) or as other non-standardized information in computer
349 documents (such as PDF-based radiology reports). Free text may be used to further characterize
350 exposure and outcome (e.g., review of patient profiles) in EHR-based data sources. To enhance
351 the efficiency of data abstraction, a range of approaches, including both existing and emerging
352 technologies (e.g., natural language processing, computer vision for images or laboratory
353 results evaluation) are increasingly being used to convert unstructured data into a computable,
354 structured data format.

355 When making secondary use of EHR data from multiple medical institutions, any differences
356 in components and format of these data, including codes used (such as disease names, drug
357 names, and laboratory test items) should be harmonised and the approach documented in the
358 protocol. EHR data typically capture the medical encounter with the health care provider but
359 may not reflect the actual delivery of healthcare (e.g., medicines that are ordered but not
360 dispensed or administered) and may require additional linkage (e.g., to pharmacy records). In
361 addition, obtaining comprehensive history of medicine use, or medical care data on patients

362 with certain types of privacy concerns (e.g., sexually transmitted infection, substance use
 363 disorders, mental health conditions) can be challenging. Nevertheless, failure to capture these
 364 data can result in inaccurate or incomplete data.

365 **Claims Data**

366 Healthcare claims databases are often large and capture healthcare services for all individuals
 367 covered by a health insurance program(s). Typically, once claims for all healthcare provided to
 368 individuals within a health insurance program are fully adjudicated (i.e., final payment
 369 decisions made by insurance companies or claims processors), they are aggregated into a
 370 database that reflects a more complete view of services. Some databases will contain a mix of
 371 open (in-process) and closed (paid/denied) claims and the researcher should understand the
 372 dynamic nature of the data in these cases. Without linkage to other data sources, it is often not
 373 possible to obtain information about healthcare visits, results from laboratory testing, outcomes
 374 of offspring in pregnancy studies, many vaccinations, injuries from accidents, and other care
 375 not covered by health insurance. These issues may be due to numerous factors, including health
 376 insurance coverage policies and seeking medical care outside of the insurance system (e.g.,
 377 self-paid/self-care treatments, procedures insured by worker's accident insurance, and motor
 378 vehicle liability insurance).

379 **Registries**

380 A *registry* is an organized system that collects prespecified uniform data from a population
 381 defined by a specific disease, condition, or exposure. Registries may be further described as
 382 “Patient Registries” or “Product Registries” to indicate defining characteristics for registry
 383 entry. The former highlights a focus on collecting data from patients with a certain disease,
 384 specific populations, such as pregnant or lactating people, or individuals with a specific
 385 condition, such as a birth defect or a molecular or genomic feature. The latter is a system by
 386 which sponsors collect data on patients exposed to a specific health care product or class of
 387 products.

388 An already established registry may be used to collect data for reasons other than originally
 389 intended. If a study makes secondary use of registry data, the same considerations and fit-for-
 390 purpose assessment relevant to secondary sources such as EHR and claims data should be
 391 applied to evaluate the suitability of the registry to answer the research question, e.g., taking
 392 into account the registry population, data elements collected, including longitudinally,
 393 frequency of data assessments, and calendar time, level of data quality, and governance

394 (including aspects on data sharing and data access). Additional considerations may include the
395 type of registry and the impact of methods involved in patient selection on the
396 representativeness of the population relative to the target population (such as geographic
397 factors, total number of patients in the registry, number eligible, number of new patients
398 entering the registry per year and number lost per year with reasons for exit). If data necessary
399 to answer the research question(s) are not routinely collected within established registries,
400 linkage to external data sources or supplemental data collection through other means should be
401 explored. In some cases, de novo registries or other study designs may be required (e.g., need
402 for an adequate comparator population in a single-arm product registry, key measures of
403 exposure or covariates such as duration, dose and route of therapy administration, or intractable
404 channeling bias requiring randomization).

405 **Data Collected by Digital Health Technologies (DHT)**

406 *Digital health technologies (DHTs)* are systems that use computing platforms, connectivity,
407 software, and/or sensors for health care and related uses. These technologies span a wide range
408 of uses, from applications in general wellness to applications as a medical device. They include
409 technologies intended for use as a medical product, in a medical product, or as an adjunct to
410 other medical products (devices, drugs, and biologics). They may also be used to develop or
411 study medical products. Technological advances have increased the range of data sources that
412 can be used to complement traditional ones and may provide insights into or relevant to safety
413 (and effectiveness) of health interventions. These technologies should be subject to the same
414 fit-for-purpose assessments as other data sources. There may be a need to specify DHTs (e.g.,
415 version, software, hardware, manufacturer), or to harmonize data across different types of
416 devices. Depending on the data source maturity, greater validation work may be needed.

417 **Federated Data Networks**

418 *Federated Data Networks (FDNs)* enable distributed analyses combining data or results across
419 multiple databases. When choosing to use a FDN for a study, there are specific issues unique
420 to these systems that should be considered, such as the FDN's transformation of data into
421 common data models (CDMs), and the differences between systems from which the data arise.
422 Governance of federated networks (centralized or decentralized) also needs to be taken into
423 account, as it has an impact on the operational aspects of a study.

424 Under the FDN framework, different approaches can be applied for combining data or results
425 from multiple databases. A common characteristic of all approaches is the fact that data partners

426 maintain physical and operational control over electronic data in their existing environment
427 and therefore the data extraction is always done locally. Differences, however, exist in the
428 following areas: use of a common protocol; use of a CDM; and where and how the data analysis
429 is conducted.

430 The choice of data captured in a CDM is optimized for the types of data measures and detail
431 needed for the intended use. Therefore, data in CDM-driven networks rarely contain all of the
432 source information present at the individual databases, and the data elements chosen for a given
433 CDM network may not be sufficient for all research purposes or questions.

434 FDNs can provide unique advantages that can assist with addressing drug safety questions,
435 such as:

- 436 • Decreasing the time to conduct a study, either through pre-developed analyses, or by
437 increasing the size of study populations as this shortens the time needed to obtain the
438 desired sample size. Large sample sizes may facilitate research on rare events, rare
439 diseases, and less common drug exposures;
- 440 • Multi-database studies may provide additional knowledge on whether a drug safety
441 issue exists in different populations or across countries and thereby may reveal causes
442 of differential drug effects, inform the generalizability of results, the consistency of
443 information and the impact of biases on estimates;
- 444 • Heterogeneity of treatment options and utilization patterns across institutions,
445 communities or countries may allow for a more complete understanding of the effect
446 of individual medicines; and
- 447 • Involvement of experts from various countries addressing terminologies, coding in
448 databases and research practices provides opportunities to increase consistency of
449 results of pharmacoepidemiological studies.

450 **Data Linkage**

451 Data linkage can be used to increase the breadth and depth of data on individual patients over
452 time and may be utilized to allow access to other data sources to support validation efforts.
453 Linkage of data sources such as cancer or mortality registries linked to claims or EHR may
454 result in a higher quality study by including data not in the original data source. It is important
455 to have a comprehensive understanding of the data and to assess the accuracy and completeness
456 of the linkage and the resulting linked data. In some circumstances, chart review or text entries
457 in electronic format linked to coded entries can be useful for exposure, outcome, and covariate

458 identification.

459 Conceptually, a data linkage may be undertaken within a database (e.g., mother–infant linkage)
460 or across databases (e.g., vital records, biobank). If the study involves a data linkage, the
461 protocol should describe each data source, the information that will be obtained, linkage
462 methods, and the accuracy and completeness of data linkages over time. If the study involves
463 generating additional data (e.g., interviews, surveys, computerized or mobile-application
464 questionnaires, measurements through digital health technologies), the protocol should
465 describe the methods of data collection and linkage, explanations of the data elements used for
466 linkage, and what will be done if imperfect linkage exists, or contradictory data are found
467 across linked data sources.

468 **5.2.3 Data Standardization**

469 Data standardization is relevant to multi-database studies, including federated data networks.
470 There are several challenges to consider in standardizing study data derived from RWD
471 sources. These challenges to standardization include but are not limited to:

- 472 • The type of information the sources contain (e.g., diagnoses, procedures, medications);
- 473 • The variety of RWD sources and the level of consistency in formats and coding
474 languages, including differences in source data captured regionally and globally using
475 different standards and terminologies;
- 476 • Differences in healthcare systems, such as business processes and local healthcare
477 practice patterns, database structure, vocabularies, coding systems, and de-
478 identification methodologies used to protect patient data when shared.

479 Coding systems for diagnoses, medicines, and laboratory data, among others, are updated
480 regularly. Therefore, plans for mapping coding systems as they evolve/change should be
481 addressed at the protocol stage. Moreover, care should be given when re-using a code list from
482 another study, as code lists reflect the individual study objectives, methods, and the time in
483 which they were created.

484 *A free-text/unstructured component exists in some databases, and can be used to define*
485 *inclusion criteria, exclusion criteria, exposures, outcomes, follow-up, and covariates. Each*
486 *free-text component may be transferred into a structured table which prompts users to specify*
487 *what is measured, the timing of measurement, the care setting, type of codes that are used to*
488 *define the measure as well as the sources for any algorithms used to derive study measures,*

489 *e.g., defining exposures, outcomes, or covariates. The process for creating a structured*
490 *variable from unstructured data should be provided in the study documentation.*

491 **5.2.4 Missing Data**

492 Missing data are data value(s) that are not captured in the data source of interest. There are two
493 scenarios where data can be missing. The first scenario is the data are intended to be collected
494 but were not collected. The second scenario is the data are not intended to be collected in the
495 data source and therefore not available. A record in EHR systems or administrative claims
496 databases is generated only if there is an interaction of the patient with the health care system.
497 Lack of information such as a laboratory result or prescription, could be caused by different
498 circumstances, such as (1) it might not have been ordered by the health care provider; (2) it
499 may have been ordered but not conducted; (3) or it may have been conducted, but the result
500 (test, dispensing) is not recorded; or (4) there is evidence of the healthcare interaction and the
501 result was stored in the data source, but data were not in an accessible format, or lost in the
502 transformation and curation process when the final study-specific dataset was generated.
503 Approaches to handle missing data are described in further detail in Section 7, Analysis.

504 **5.2.5 Data Quality**

505 Fundamental determinants of data quality at each step in the evidence generation process, such
506 as governance and documentation need to be addressed before finalizing the protocol.
507 Depending on the data source, pharmacoepidemiologic data may lack strict **quality control**
508 (**QC**) over the process of recording, collection, and storage. This can lead to incomplete data,
509 missing key variables, or inaccurate records. The presence of such quality defects will affect
510 subsequent **data curation**, applicability, and traceability of data.

511 Compared to Good Clinical Practice (GCP), procedures for pharmacoepidemiologic data
512 quality control and **quality assurance (QA)** follow guidances specific to
513 pharmacoepidemiologic data, and detailed quality standards to be fulfilled should be in
514 accordance with local or regional regulatory requirements (see Section 6, Data Management
515 for more details on QA and QC).

516 **5.2.6 Data Collection and Data Source Sections in the Study Protocol**

517 The protocol should describe the data source(s) used and how it/they are fit-for-use to address
518 the research question of interest. In addition, the protocol should state any coding systems used
519 for classification of the exposure and outcomes (e.g., anatomical therapeutic chemical (ATC),

520 International Classification of Disease (ICD), and any methods used for data linkage). Data
521 collection methods and procedures should be described.

522 For studies that use data from multiple data sources or study sites (e.g., federated data, meta-
523 analysis, or data pooling), researchers should describe the rationale and procedures for how
524 data from different sources can be obtained and integrated with acceptable quality, given the
525 potential for heterogeneity in population characteristics, clinical practices, and coding across
526 data sources.

527 For studies with primary data collection, the identification, processing and reporting of adverse
528 events occurring in the course of treatments should be described in the protocol, in accordance
529 with relevant jurisdictional laws and regulations (see Section 8, Reporting and Submission).

530 **5.3 Target/Study Population**

531 The target population is the population about which one wants to make an inference (e.g.,
532 children aged 12-16 with attention deficit hyperactivity disorder). The study population is
533 intended to be representative of the target population from which data will be evaluated to
534 answer the research question. The study population is defined via inclusion and exclusion
535 criteria (e.g., demographic factors, medical conditions, disease status, severity, biomarkers) and
536 identified based on the following elements, among others:

- 537 • Time points, such as index dates for inclusion in the study, defined lookback period (e.g.,
538 to identify new users);
- 539 • Key variables (see Feasibility Assessment(s)) used to select the study population and how
540 they should be validated (see Bias and Confounding); and
- 541 • The completeness and accuracy of the data elements to fulfil the inclusion and exclusion
542 criteria (see **Error! Reference source not found.**).

543 **5.4 Exposures, Outcomes, Covariates**

544 If the initial feasibility assessment has indicated that the exposures, outcomes, and covariates
545 of interest are likely to be adequately captured in the selected data sources, then defining and
546 operationalizing these elements should proceed. This process generally starts with the creation
547 of a *conceptual definition* which is initiated at the time of initial database selection. The
548 conceptual definition should reflect current medical and scientific thinking regarding the
549 variable of interest, such as: (1) clinical criteria to define a condition for population selection
550 or as an outcome of interest or a covariate; or (2) measurement of drug intake to define an

551 exposure of interest. The conceptual definition should include a detailed description of the data
552 elements that would characterize the exposure, outcome, or covariates.

553 Utilizing the key data elements identified during the feasibility phase, this conceptual definition
554 is then developed into the *operational definition*. An operational definition should be
555 developed based on the conceptual definition to extract the most complete and accurate data
556 from the data source. In many studies using EHRs or claims data, the operational definition
557 will be a code-based electronic algorithm using structured data elements. In other studies, the
558 operational definition may be derived from extracting relevant information from unstructured
559 data or constructing an algorithm that combines structured and unstructured data elements.
560 Operational definitions can also specify additional data collection, such as a patient survey,
561 when appropriate. Researchers should consider the following areas when developing exposure,
562 outcome, and covariate definition(s):

- 563 • Whether it is possible to translate a conceptual definition of the exposures, outcomes,
564 and covariates into one that can be operationalized in selected data source(s);
- 565 • Whether the operational definition adequately captures all elements of the conceptual
566 definition; and
- 567 • Whether the operational definitions and the performance characteristics are adequate in
568 the chosen data source(s) based on the research question (see Validation).

569 The conceptual definition is often referred to as the *phenotype*. The protocol should include a
570 detailed description of the operational definition, sometimes referred to as the computable
571 phenotype (including the coding system and rationale) and the associated limitations (e.g.,
572 measurement bias, proxies), the potential impact of misclassification, and how these limitations
573 can be mitigated through the study design and analysis. For unstructured data, a detailed
574 description, rationale for use, search criteria to identify outcomes/exposures/covariates, and the
575 list of codes or concepts should be provided. The operational definitions should be documented
576 in the protocol and/or the statistical analysis plan.

577 Operational definitions developed for one data source or study population might perform
578 differently in other sources or populations in terms of sensitivity and specificity due to
579 database-specific characteristics as well as variations in the disease epidemiology across
580 populations and databases. If utilizing or adapting a definition used or validated in other studies
581 or databases, applicability must be justified.

582 When identifying exposures and outcomes in a database for a specific study, data related to

583 these types of information are usually coded. When selecting a data source, appropriateness of
584 the coding system for defining the exposures and outcomes should be confirmed.

585 The following elements require consideration during protocol development and are described
586 in more detail below:

- 587 • Data source/type and structure;
- 588 • Development of exposure, outcome and covariate definitions and the method used to
589 identify them;
- 590 • Development and performance of the operational definitions, including time points, data
591 types (structured, unstructured), variable types (categorical, continuous), transformation of
592 variable types, code types, mapping of dictionary codes (e.g., ICD-10 to MedDRA) when
593 applicable, and the mechanism of evaluation (selection of gold standard) and performance
594 measures;
- 595 • Mapping of the available data elements against those required for the research question;
- 596 • Documentation of variable validity and appropriateness of applying previously used
597 algorithms to the database/population of interest; and
- 598 • Potential impact of misclassification on study validity and interpretation.

599 **5.4.1 Exposure**

600 **Conceptual Definition**

601 An exposure is the medicine of interest (and dosing or regimen) being evaluated in the proposed
602 study. The product of interest is referred to as the treatment and may be compared to no
603 treatment, standard of care, another treatment, or a combination of the above.

604 Exposure definitions can have differing levels of granularity, such as ever exposed vs. never
605 exposed, duration of exposure, user type (e.g., incident vs. prevalent), exposure windows (e.g.,
606 current vs. past exposure), also referred to as risk periods or risk windows, multiple exposure
607 (e.g., use of more than one medicine or concomitant vaccinations), treatment switching,
608 sequencing (e.g., first line or second line) or dosage (e.g., current dosage, cumulative dosage
609 over time). Consideration should be given to both the requirements of the study design and the
610 availability of data. The exposure definition should include information about the medicine
611 dose, brand, formulation, strength, route, timing, frequency, and duration (as applicable). It
612 may also be necessary to describe the manufacturer as part of the product identification (e.g.,
613 for a medicine with the same active substance name made by different manufacturers). This

614 may require an understanding of the pharmacological or biological properties of the medicine,
615 or members of the product class.

616 **Operationalizing Exposure**

617 *By Medicine Type, Route, and Setting*

618 When translating the conceptual definition to the operational definition, there are uncertainties
619 that should be considered, and justified in a discussion of strengths and limitations in the
620 protocol. For example:

- 621 • Medicines that are prescribed are not necessarily dispensed;
- 622 • Medicines that are dispensed are not necessarily used or administered;
- 623 • Patient compliance and ability to provide an accurate account of intake;
- 624 • Exposures that are not captured in the data source such as samples, low-cost medicines,
625 non-prescription medicines, and immunizations offered in the workplace; and
- 626 • Coding systems used to identify exposures (e.g., NDC, RxNorm, HCPCS).

627 The setting of administration should be considered carefully. Infusions may be administered in
628 private clinics or on an outpatient basis (e.g., home care) in addition to in-hospital, so setting
629 and treatment patterns should be considered in terms of potential requirements for data
630 linkages.

631 For vaccines, it is essential to have granular information on brand, dose schedule,
632 coadministration with other vaccines, and sometimes batch number or administration route and
633 site. These data may require linkage to vaccination registries.

634 *By Medicine Dose, Timing, and Duration of Exposure*

635 The exposure (i.e., dose, dosage regimen) to the medicine of interest should be well-defined
636 and measured. Consideration should be given to the timing of exposure for medicines (e.g., the
637 relevant exposure window, relative to the onset of the outcome), and this may be especially
638 difficult when “as needed” or non-prescription medicines are an exposure of interest. When
639 defining the exposure period, it is necessary to decide whether the start date of exposure (the
640 index date) is the date of prescription, the date of dispensing, or the date of administration.
641 Because patients may not refill their prescriptions exactly on time or, alternatively, may refill
642 their prescriptions early, gaps or stockpiling in therapy may exist and may be reflected in the
643 data. Allowable gaps between dispensing to construct exposure episodes and the gaps between
644 exposure episodes should be considered when determining whether an exposure period is

645 continuous. Conditions for the completion of an exposure period should also be considered and
646 explicitly defined (e.g., no record of a new prescription in preceding six months), noting
647 limitations such as the potential of a drug being prescribed to a patient in another setting that
648 may not be captured in the dataset used for the study.

649 **5.4.2 Outcome**

650 **Conceptual Definition**

651 Defining an outcome of interest should be based on the clinical, biological, psychological, and
652 functional concepts of the condition, as appropriate. This conceptual definition should reflect
653 the medical and scientific understanding of the condition. Considerations for how outcomes
654 should be identified will include whether cases can be identified as true incident (vs. prevalent),
655 the latency, and whether the outcome presents with exacerbations or as recurrent episodes. The
656 definition should include a detailed description of the data elements that would confirm the
657 outcome (e.g., signs, symptoms, laboratory and radiology results).

658 Clinical outcome definitions should contain diagnostic criteria, measuring methods and their
659 quality control (if any), measurement tools (e.g., the use of questionnaire scales), calculation
660 methods, measurement time points, variable types, transformation of variable types (e.g., from
661 quantitative to qualitative variable), and mechanism of endpoint event evaluation (e.g., the
662 operation mechanism of the endpoint event committee).

663 **Operationalizing Outcome**

664 An operational definition is one that can be implemented independently using the data available
665 in the proposed study with acceptable performance to meet the goals of the study. The
666 conceptual definition is operationalized using diagnosis and procedure codes (e.g., ICD-9-CM,
667 ICD-10, Read, MedDRA), laboratory tests (e.g., Logical Observation Identifiers Names and
668 Codes [LOINC]) and values, unstructured data (e.g., physician's encounter notes, radiology, or
669 pathology reports), or measurement tools such as questionnaire scales. Consideration for
670 changes in coding or the underlying EHR systems over time is essential. If unstructured data
671 are used, detailed description and rationale for the methods and tools utilized and validation of
672 those methods should be provided.

673 Single appearances of a diagnosis code may indicate a rule out diagnosis or lack adequate
674 specificity. Instead, consider whether a valid definition of outcome can be achieved by
675 combining medicines, laboratory data, and medical procedures used for diagnosis or treatment,

676 rather than operationalizing the outcome only based on the disease or diagnosis (e.g., a
677 thromboembolism diagnosis code plus treatment with anticoagulant, anaphylaxis code plus use
678 of epinephrine). In some studies, in which the outcome is complex to define, information on
679 the specialty of the physician making the diagnosis might help provide additional reassurance
680 regarding the quality of the information used to determine the outcome. Mortality as an
681 outcome may not be included in electronic health care data unless the patient was under medical
682 care at time of death. Linkage to external vital statistics resources may be necessary.

683 When considering use of previously developed operational definitions, researchers should
684 consider secular trends in disease, diagnosis, or changes in coding practices that may
685 necessitate assessment using more recent data. Published *case definitions* of outcomes may be
686 used but are not necessarily compatible with the information available in a given RWD data
687 set. When proposing to use an operational definition that has been assessed in a prior study,
688 ideally select those assessed in the same data source and in a similar study population. In
689 addition, the quality of prior studies to establish sensitivity, specificity, and predictive values
690 should always be evaluated. Applicability of a definition used in a prior study or validated in
691 another data source will depend on an assessment of its external validity with a justification
692 provided in the protocol.

693 When conducting a study using data from multiple data sources (databases, institutions, sites),
694 define the outcome considering the data differences between sources, such as diagnosis coding,
695 laboratory reference ranges and medication records. A complete understanding of the timing
696 and relationship between these elements is essential. For example, there are situations where
697 the start date of treatment on the claims data and the date of diagnosis on the EHR data may
698 not match for the same patient.

699 When outcomes to measure patient experience are included (e.g., quality of life, subjective
700 severity of disease), the protocol should specify how the outcome measure is defined,
701 constructed, and validated, and the procedures for data collection. The reason for the data
702 collection and the nature of the healthcare system that generated the data should also be
703 described as they can impact the quality of the available information and the presence of
704 potential biases.

705 **5.4.3 Covariates**

706 Covariates are variables that are neither an exposure nor an outcome of interest, but instead are
707 measured because they either characterize a population or are a potential confounder or effect

708 modifier to account for in study design or analysis. As with exposure and outcome, the
 709 definition moves from a conceptual definition to an operational definition based upon clinical,
 710 biological, psychological, and functional concepts, as appropriate. The definition should
 711 include a detailed description of the data elements used to construct the covariate.

712 The potential for confounding and effect measure modification should be considered and
 713 planned for during protocol development. For example, the potential for *effect modification*
 714 by demographic variables (e.g., age, sex, race, ethnicity), other exposures (e.g., biologically
 715 active herbals) or pertinent comorbidities should be documented in the study, and relevant
 716 effect modifiers should be available in the chosen data source.

- 717 • **Confounding:** Confounding occurs when the estimate of measure of association is
 718 distorted by the presence of another factor. For a variable to be a confounder, it must
 719 be associated with both the exposure and the outcome, without being in the causal
 720 pathway.
- 721 • **Effect Modification:** Effect modification occurs when the effect of a single exposure
 722 on an outcome depends on the values of another variable, i.e., the effect modifier, which
 723 does not necessarily need to be involved in the causal pathway.

724 **Conceptual Definition**

725 Definitions of covariates needed in a study should be identified and a determination made on
 726 whether it can be directly operationalized in a given data source. When the covariate is not
 727 available in the chosen data source, researchers may consider whether proxies for the covariate
 728 are appropriate.

729 **Operationalizing Covariates**

730 Moving from a conceptual to operational definition proceeds as with exposure and outcome.
 731 Covariates may be used to characterize cohorts, to develop propensity scores, to stratify or
 732 match patients, evaluate effect modification and adjust for confounding. Covariates are
 733 typically identified and assessed during the period before the start of the exposure of interest
 734 (baseline). Assessment of baseline covariates can be performed using different periods of time.
 735 The length of this lookback period is selected by considering factors such as changes in coding
 736 or medical practice, expected frequency of medical encounters, relevance to the research
 737 question, and the impact on study power. Covariates and may also be assessed during the
 738 observation period, either as static or time varying variables. Reliable assessment of covariates
 739 is therefore essential for the validity of results, including the timing of assessments for each of

740 the covariates. A given database may or may not be suitable for studying a research question
741 depending on the availability of information on these covariates. Researchers should provide
742 the developed operational definition, including codes and settings of care, for all covariates in
743 the protocol.

744 **5.5 Bias and Confounding**

745 To obtain a valid and precise estimate of the effect of exposure on the outcome of interest,
746 studies must address two sources of error. Unlike random error, systematic error (bias,
747 confounding) cannot be addressed by increasing sample size. Rather, it is typically addressed
748 in the design, conduct, and analysis stages. From the epidemiological standpoint, it is useful to
749 differentiate the concepts of bias (e.g., selection bias, information bias, resulting from design
750 or measurement errors) and confounding because they arise from distinct mechanisms and may
751 be addressed by distinct methods and approaches in study design and analysis. The design and
752 analysis stages should include evaluation of any potential biases such as information bias and
753 selection bias which can be due to the inclusion/exclusion criteria or loss to follow-up, as well
754 as evaluation of any confounding that may arise, especially if some data elements cannot be
755 collected or measured. Therefore, the handling of missing data should also be prespecified in
756 the Data Management section (see Section 6, Data Management) or Analysis section (see
757 Section 7.1, Statistical Analysis) of the protocol.

758 The proposed data source should be evaluated to determine whether it is adequate to capture
759 information on important factors so that bias and confounding may be adequately controlled.
760 Linkage with other data sources or additional data collection to expand the capture of important
761 variables that are unmeasured or imperfectly measured in the original data source should be
762 considered. Sources of bias and confounding should be considered, and decisions to address
763 should be justified during the design stage with a plan to evaluate the influence of bias and
764 confounding; these should be included in the protocol, analysis plan or final report.

765 **5.5.1 Selection Bias**

766 There are different types of selection bias such as referral bias, self-selection bias, prevalent
767 user bias, and loss to follow-up (time-related bias). Different forms of selection bias may be
768 addressed in either the design (preferred) or analysis stages.

769 A common type of selection bias is prevalent user bias, which can arise when prevalent users
770 of a medicine are included in a pharmacoepidemiologic study, i.e. patients already taking a

771 therapy for some time before study follow-up began. Prevalent users are ‘survivors’ of the early
772 period of pharmacotherapy that is not captured in the study. This can introduce selection bias
773 if the risk varies with time. For example, subjects who initiated a new medicine, experienced a
774 safety event, and then discontinued the medicine may not be included in the study, thereby
775 leading to a potential underestimation of the risk among the treated.

776 **5.5.2 Information Bias**

777 Information bias arises when misclassification of binary or categorical variables or
778 mismeasurement of continuous variables exists. Whereas internal validity should always be
779 optimized, and misclassification of key variables should be minimized to accurately estimate
780 the effect of exposure on the outcome, some degree of misclassification may be acceptable in
781 some studies depending on the study question and regulatory purpose and should be determined
782 on a case-by-case basis. Overall, the extent of variable validation (see [Validation](#)) should be
783 determined by the necessary level of certainty and the implication of potential misclassification
784 on study inference. As discussed in [Section 3](#), a plan to use *quantitative bias analysis* may be
785 useful when evaluating the direction and magnitude of biases to inform strategies for bias
786 mitigation, and how the study biases may influence the interpretation of the study (see
787 [Analysis](#)).

788 **5.5.3 Immortal Time Bias**

789 Immortal time bias refers to a period of cohort follow-up time during which an outcome of
790 interest cannot occur. Selection of an appropriate index date is essential to avoid the risk of
791 immortal time bias and other time-related biases. This risk may be mitigated by design
792 frameworks (see [Research Question](#)), as this approach aligns assessment of eligibility and
793 baseline information with start of follow-up [4].

794 **5.5.4 Confounding**

795 Researchers are typically unable to capture all potential confounders that are relevant to a
796 research question, introducing the potential for unmeasured or residual confounding. In
797 pharmacoepidemiology, commonly considered confounding factors include demographics,
798 indication for treatment, disease severity, previous medication use, and comedications,
799 comorbidities, prognostic characteristics, frailty, and others, depending on the study question.
800 A number of approaches are available to address or evaluate unmeasured confounding,
801 including high dimensional propensity scores, negative controls, and linkages to external data

802 sources such as surveys that include data on confounders unmeasured in the study database.
803 The presence and impact of potential confounding factors should be considered in the study
804 design phase. Directed acyclic graphs can be used to understand the relations between the
805 variables and identify potential confounding and intermediate effects in a longitudinal study,
806 and the impacts of these assessed using quantitative bias analysis, as discussed in the Analysis
807 section (see [Analysis](#)).

808 **5.6 Validation**

809 Validity is the extent to which a concept (variable) is accurately measured in a study by the
810 operational definition. Validation of exposure, outcome, and key covariates is important for
811 internal validity of pharmacoepidemiological studies. There are various approaches to
812 validation, which may be data-source specific. These may include complete verification, partial
813 verification, clinical expert review, review of patient claims, or profile history. Validation
814 efforts should be commensurate with the level of evidence required, such as validating the
815 outcome variable for all potential cases or non-cases or verifying the performance of an
816 operational definition to identify cases and non-cases. For databases routinely used in research,
817 documented validation of key variables may have been done previously. Any extrapolation of
818 a previous validation study should however consider the effect of any differences in prevalence
819 and inclusion and exclusion criteria, the distribution and analysis of risk factors as well as
820 subsequent changes to health care, procedures, and coding. Sponsors should have early
821 interactions with regulatory authorities to discuss and agree upon a proposed validation
822 approach, such as partial vs. full, or adoption of definitions validated previously. A justification
823 for validation approach should include the data source, population, time frame, performance,
824 reference standard, and a discussion of the applicability of the proposed operational definitions
825 considering the level of evidence required.

826 Validation studies should be conducted under a separate protocol. When validating an
827 operational definition, prespecify the metrics to be reported (e.g., sensitivity, specificity,
828 positive predictive value, negative predictive value, kappa statistic), and describe how they will
829 be measured. The trade-off between false-positive and false-negative cases should be
830 considered when selecting an operational definition and identifying the proper validation
831 approach to support internal validity. For instance, when cases are rare, one may need to select
832 a highly sensitive operational definition and then validate all potential cases.

833 If several operational definitions are under consideration, the performance of each should be

834 evaluated and quantified using bias analysis in the design stage. This is distinct from common
835 sensitivity analyses conducted in the analysis stage.

836 **6 Data Management**

837 Management of data quality for a pharmacoepidemiological study depends on various factors,
838 including the source of the data and the planned use of the study results. A data management
839 and/or data curation plan should be developed prior to study initiation. The quality assurance
840 and quality control (QA/QC) plan should be developed before an analysis is undertaken and
841 the various factors (e.g., data management by the data holder, quality defects of the data,
842 inadequate data processing and analysis, or inadequate training) influencing quality should be
843 identified and addressed to preserve the integrity of the study. Detailed quality standards to be
844 fulfilled should be in accordance with local or regional regulatory requirements.

845 To facilitate regulatory review, where submission of datasets is a regulatory requirement, a
846 description of the context, content, structure of files, and steps used to create the files should
847 be included. Datasets must be retained in accordance with relevant regulatory requirements in
848 the region(s) to which they will be submitted. Any migration of data and documents to new
849 media or a new format should be verified to ensure long-term readability and to maintain
850 integrity.

851 **Data Management Plan**

852 Data quality assurance processes, policies, and procedures should account for potential risks to
853 data quality, including errors in interpretation or coding; errors in data entry, transfer, or
854 transformation accuracy; errors of intent; inadequate training; *data completeness*; and *data*
855 *consistency*.

856 A description of data storage, management, and statistical software should be included in the
857 protocol. All procedures used to obtain, verify, and promote the integrity of the analytic dataset
858 should be recorded in sufficient detail so that they can be replicated. Data security should
859 always be maintained by limiting access to authorized individuals.

860 **Quality Assurance and Quality Control**

861 **6.1 Data Holder**

862 QA and QC procedures used by the *data holders* may include ensuring reliability of data
863 collection and management; the frequency and type of any data error corrections or changes in

864 data adjudication policies implemented by the data holders during the relevant period of data
865 collection; peer-reviewed publications examining data quality and/or validity, updates and
866 changes in coding practices (e.g., International Classification of Diseases, ICD codes) across
867 the study period that are relevant to the outcomes of interest; changes in key data elements
868 during the study time frame and their potential effect on the study; the extent of missing data
869 over time (i.e., the percentage of data not available for a particular variable of interest), and
870 procedures (e.g., exclusion, imputation) employed to handle these issues.

871 **6.2 Researchers**

872 While the data holder maintains control of the data and is responsible for the underlying data
873 quality, the researchers are responsible for aligning QC and QA procedures with data holders
874 to ensure transparency, understanding of data strengths/limitations, and meeting the standards
875 of quality criteria required by the regulatory authorities. Further, the researcher is responsible
876 for the management and quality assurance of all data cleaning, processing, and analytic
877 datasets. To balance the need for sufficient quality assurance with reasonable resource
878 expenditure for a particular purpose, a risk-based approach to quality assurance is
879 recommended. Issues that are essential to determining the reliability and relevance of the data
880 should be addressed in the protocol, and include QA/QC procedures for data accrual, curation,
881 and transformation into the final study-specific dataset.

882 The researcher is responsible for implementing and maintaining QA/QC systems with written
883 procedures. This is to ensure that studies are conducted, and results are generated, documented,
884 and reported in compliance with the protocol, regional laws, ethical considerations, and the
885 applicable regulatory requirement(s). Documentation of these processes may include, but are
886 not limited to electronic documentation (i.e., metadata-driven audit trails, quality control
887 procedures) of data additions, deletions, or alterations from the data source to the final study
888 analytic dataset(s). Researchers should also document changes to data and the potential impacts
889 of these changes for conducting this specific study. Methods for quality assurance and quality
890 control of analytic programming should be described in the protocol, such as the process to
891 inspect code and/or replicate code, or whether an analytical code that was previously
892 QA/QC'ed is being used.

893 **7 Analysis**

894 The analytic strategy includes descriptive and inferential analyses to address the study

895 objectives, while accounting for potential sources of bias and confounding. In addition, the
 896 strategy should also include an empirical evaluation of unmeasured or mismeasured
 897 confounding and other sources of bias. The statistical analysis should be prespecified, reflect
 898 the information gained from the feasibility assessments(s), and be developed to meet the study
 899 objectives. An overview of the *statistical analysis plan (SAP)* should be provided in the
 900 protocol. The complete SAP should be provided as a standalone document, or as a detailed
 901 section of the protocol. It is recommended that the approach chosen should be discussed with
 902 health authorities, keeping in mind that the protocol and SAP are highly interdependent. The
 903 SAP should provide sufficient detail to allow replication of the study to help ensure confidence
 904 in the results.

905 In some studies, data driven analyses may be performed and it is important to distinguish
 906 between those that are pre-specified and those that are *post-hoc*. Pre-specified analyses such as
 907 those used for covariate selection should be documented in the protocol and analysis plan and
 908 deviations from the plan documented in the final report. Post-hoc analyses are often conducted
 909 in response to observations in the data to help in the interpretation of results and should be
 910 described and justified in the final report.

911 Researchers should consider developing a timeline of the analyses that will be performed
 912 during the conduct of the study (e.g., accrual, descriptive analyses, inferential analyses,
 913 sensitivity analyses, and quantitative bias analysis).

914 Proactive planning is required when conducting a multi database study or when using an FDN,
 915 as the analytic strategy is impacted by the types of databases or the FDN under consideration.
 916 Specific issues may need to be considered in the analysis plan, such as the independent related
 917 analyses performed in each data source or FDN which may require meta-analytic techniques.

918 **7.1 Statistical Analysis**

919 **7.1.1 Primary Analyses**

920 The analysis should be directed towards the unbiased estimation of the epidemiological
 921 parameters of interest (e.g., risk or rate differences and risk or rate ratios). The analysis section
 922 is where a description and justification for the chosen approaches for the statistical analyses
 923 should be described, including the assumptions and conditions.

924 The following aspects and elements may be considered for inclusion: descriptive analyses,
 925 subgroup analyses, methods of estimation and associated assumptions needed for analysis,

926 estimate of the anticipated study size/power/statistical precision, plans to control for
927 confounding and bias (e.g., misclassification, selection bias, information bias, time-related
928 bias, and impact on validity of results), assessment of population comparability, sensitivity
929 analyses, type I error control (e.g., for sequential analysis), assessment of representativeness
930 and plans for handling missing data.

931 If the analysis proposes to use machine learning or other derivation methods, specify the
932 assumptions and parameters of the computer algorithms used, the data source from which the
933 information was used to build the algorithm, whether the algorithm was supervised (i.e., using
934 input and review by experts) or unsupervised, and the metrics associated with validation of the
935 methods.

936 **7.1.2 Missing Data**

937 Researchers should develop the protocol and the statistical analysis plan with an understanding
938 of reasons for the presence and absence of information in the underlying data; consider data
939 linkage and or imputation to address missing data, and address the implications of the extent
940 of missing data on study findings (see Bias and Confounding). Descriptive analyses should be
941 included to characterize missing data. Assumptions regarding the missing data (e.g., missing at
942 random, missing not at random) underlying the statistical analysis for study outcomes and
943 important covariates should be supported and the implications of missing data considered. The
944 analysis should address missing data in line with the methods described in the statistical
945 analysis plan. The extent and implications of missing data on study findings should be
946 described.

947 **7.1.3 Sensitivity Analyses**

948 When planning for sensitivity analyses, a rationale for each analysis should be provided with
949 the strengths and limitations of each analysis. Sensitivity analyses should be conducted to
950 examine the effect of varying potentially critical assumptions of the analysis, such as those
951 relating to design, estimands, exposure definition, outcome definition, missing data, and
952 limitations of the data source(s) selected. The analyses can facilitate better interpretation of
953 study results in light of the extent of uncertainty noted.

954 Quantitative bias analysis evaluates the impact of potential bias on the measure of association.
955 The protocol should pre-specify the indices (e.g., sensitivity, specificity, positive [PPV] and
956 negative [NPV] predictive values) that will be used for quantifying bias and describe how the

957 selected indices will be measured when validating variables of interest. The precision of the
958 bias-adjusted effect estimates should be quantified using confidence intervals. These analyses
959 may facilitate interpretation of study results.

960 **8 Reporting and Submission**

961 **8.1 Reporting of Adverse Events, Adverse Drug Reactions, and Product Quality**

962 **Complaints**

963 Adverse events, adverse drug reactions, and product quality complaints identified during the
964 conduct of a study may require reporting to the relevant regulatory authority. Reporting
965 requirements may vary by party (e.g., marketing authorization holder (MAH), other sponsor or
966 applicant, investigator, or independent research group) and by region, due to differences in
967 regulatory reporting requirements. The ICH E2D guideline on Post Approval Safety Data
968 Management provides guidance for MAHs on reporting of individual case safety reports for
969 adverse events and adverse drug reactions. For other reporting requirements (and for parties
970 outside the scope of ICH E2D), refer to applicable laws and regulations and, as appropriate,
971 pharmacovigilance guidelines.

972 **8.2 Formatting and Content of Study Documents for Submission to Regulatory**

973 **Authorities**

974 Sponsors should discuss with regulators the required study documents and timetables for
975 submission. These documents may vary by regulator, can include the feasibility assessments,
976 protocol, analysis plan, and interim and final reports. In the absence of specific formatting and
977 content required by regulators, sponsors may utilize frameworks developed by the scientific
978 community as a guide for document development, such as ISPE/ISPOR's HARmonized
979 Protocol Template to Enhance Reproducibility (HARPER) [1, 5].

980 **9 Dissemination and Communication of Study Materials and Findings**

981 For transparency, to support scientific exchange, and to allow the conduct of reproducible
982 research, even where not mandated by regulatory requirements, it is encouraged that
983 researchers make study materials publicly available. It is encouraged that the protocol and
984 statistical analysis plan be made publicly available in appropriate registers before study
985 initiation, and study reports upon completion. Further vehicles for dissemination and
986 communication of study results may include non-regulatory submission in scientific fora,

987 scientific publications, and patient or practitioner-focused communications.

988 Several guidelines exist that provide recommendations for reporting studies in medical
 989 literature. These include The Reporting of studies Conducted using Observational Routinely
 990 collected health Data (RECORD) Statement, RECORD-PE, and “Recommendations for the
 991 Conduct, Reporting, Editing, and Publication of Scholarly work in Medical Journals,”
 992 established by the International Committee of Medical Journal Editors (ICMJE). In addition,
 993 when publishing the contents of the study, the contents of the report should be summarized so
 994 that the publication is consistent with the report. To avoid publication bias, it is recommended
 995 that the results be published even if negative or inconclusive study results are obtained with
 996 respect to study objectives or hypotheses.

997 Results of the research should be communicated to the study participants (for example, when
 998 primary data collection is used), the public, and patients, so that they may be aware of and
 999 understand the study results and their implications. Communications should include a factual
 1000 summary of the overall study results in an objective, balanced and nonpromotional manner,
 1001 including relevant safety information and any limitations of the study.

1002 **10 Study Documentation and Record Retention**

1003 Key documents and records related to the planning, conduct and results of a study should be
 1004 kept in compliance with applicable standards and jurisdictional requirements. Key principles
 1005 for studies utilizing RWD in post-marketing safety studies are similar to those for GCP
 1006 (especially for primary data) and Good Pharmacoepidemiological Practice (especially for
 1007 *secondary use of data*).

- 1008 • All study information, documents and records, should be recorded, handled, stored and
 1009 archived in a way that allows its accurate reporting, interpretation, verification, and that
 1010 ensures confidentiality and patient privacy in compliance with applicable privacy laws;
- 1011 • Systems are in place to ensure completeness of the study documentation, to prevent
 1012 accidental or premature loss, prevent unauthorized access, alteration, destruction,
 1013 disclosure or dissemination; and ensure that an audit trail is maintained;
- 1014 • Needed systems are in place with procedures that assure the quality of every aspect of
 1015 the documentation of study development, conduct, and reporting;
- 1016 • Study information should be readily available and directly accessible upon request by
 1017 regulatory authorities (e.g., internal or regulatory inspection ready) with risk-based

- 1018 quality checks or review processes to ensure that the primary record system is being
 1019 maintained up-to-date and that all key documents are appropriately filed; and
 1020 • All information retained at least for the duration of time required by applicable
 1021 regulatory requirements.

1022 **11 Considerations in Specific Populations**

1023 Specific (special) populations are often not enrolled in pre-approval clinical studies and include
 1024 pregnant and lactating people, infants, children, adolescents/young adults, older adults,
 1025 immunocompromised patients, and people with disabilities and/or rare disorders. Therefore,
 1026 post-market pharmacoepidemiological studies may provide valuable information supporting
 1027 the benefit/risk assessment of medicines in these populations. Studies in these populations may
 1028 require unique considerations when defining the study population, in addition to other
 1029 considerations applying to any studies (such as definition of exposure, confounders and
 1030 outcomes). Examples of challenges include low sample sizes for rare diseases; multiple
 1031 comorbidities and polypharmacy for older adults; and difficulty in identifying cases or disease
 1032 characteristics (e.g., duration and severity) in immunocompromised patients.

1033 **11.1 Pregnancy Studies**

1034 Specific challenges of secondary use of data in pregnancy studies include identification of
 1035 pregnancies, complexity of outcomes, and need for validated algorithms to identify gestational
 1036 age or date of conception, and maternal and infant outcomes. These challenges may necessitate
 1037 linkage within the data source (e.g., mother-child link) or complementary data sources such as
 1038 birth registries. Pregnancy registries can provide more granular clinical information on timing
 1039 of exposure, gestational age, outcomes, and covariates; however, there are challenges with such
 1040 registries, including difficulty with enrolment and retention of participants and selection bias.

1041 The dichotomous approach of ever- vs. never exposed does not reflect exposure patterns in
 1042 pregnancy and approaches to address varying risks by trimester should be considered. Attention
 1043 should be given to definition of risk windows, measurement of both conception and pregnancy
 1044 start dates, and patterns of medicine use (e.g., start and end dates, dose, frequency, duration).
 1045 A valid estimate of gestational age, from which a conception date may be estimated, is critical
 1046 for determining the timing of exposure and may require availability of linked data such as
 1047 ultrasound or laboratory data. Accurate information about gestational timing of exposure(s) can
 1048 help identify critical exposure periods and inform biological plausibility of specific effects.

1049 Exposure information in the time period just before pregnancy is often also important,
1050 especially for products with a long half-life.

1051 Outcomes include outcomes during pregnancy that affect maternal health (such as
1052 preeclampsia or gestational diabetes), spontaneous abortions, birth/neonatal outcomes, and
1053 child developmental outcomes which may extend for several years after birth. The protocol
1054 should state *a priori* criteria for defining the outcomes of interest, including their severity (e.g.,
1055 *major* birth defect), and take into account that many adverse pregnancy outcomes have
1056 substantial variation over the course of pregnancy. There are unique challenges in outcome
1057 identification for pregnancy studies and use of standard classification systems should be
1058 considered. Preterm birth and “small for gestational age” are reliably available in registries, but
1059 in administrative data may be identified through diagnostic codes or calculated using
1060 gestational age and birth weight data. Depending on the data collection methods, birth defect
1061 surveillance registries are useful as they have already been adjudicated for live births,
1062 stillbirths/fetal deaths, and elective terminations. Major congenital malformations may be
1063 recorded in the mother’s record, the infant’s record, or both.

1064 Bias and confounding in pregnancy studies include, but are not limited to, family history and
1065 confounding by indication. The analysis plan should take into account time-varying covariates
1066 in relation to the timepoint in the pregnancy.

1067 **12 Glossary**

1068 This Glossary relies on definitions sourced from ICH Guidelines, supplemented by regulatory
1069 documents, and then, relevant non-regulatory best practice documents from other sources
1070 such as professional society best practice documents and the literature.

Administrative Claims Data:

Data that arise from a person's use of the healthcare system and reimbursement of healthcare providers for that care.

(FDA, United States. Guidance Pharmacoeconomic safety studies using electronic data)

Bias:

A systematic deviation in results from the truth.

(Proposed by CIOMS Working Group X. Bias (CIOMS X: Meta-analysis 2016 | Japanese)

Case Definition:

The clinical, biological, psychological, and functional concepts of the condition, that reflect the medical and scientific understanding of the condition.

(FDA, United States. Guidance Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products)

Common Data Model

A mechanism by which raw data are standardised to a common structure, format and terminology independently from any particular study in order to allow a combined analysis across several databases/datasets. Standardisation of structure and content allows the use of standardised applications, tools and methods across the data to answer a wide range of questions

(A Common Data Model for Europe? – Why? Which? How? – workshop report EMA/614680/2018)

Conceptual Definition:

Explains a study construct (e.g., exposure, outcomes, covariates) or feature in general or qualitative terms.

(FDA, United States. Guidance Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products)

Confounding:

Confounding results from the presence of an additional factor, known as a confounder or confounding factor, that is associated with both the exposure and the outcome, and is not in the causal pathway between exposure and the outcome. Confounding distorts the observed effect estimate for the outcome and the exposure under study.

(The European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology)

Data Accuracy:

The degree of closeness of the measured value to the nominal or known true value under prescribed conditions (or as measured by a particular method).

(M10 EWG Bioanalytical Method Validation and Study Sample Analysis -- Step 4 (final); 24 May 2022)

Data Completeness:

The “presence of the necessary data” (National Institutes of Health 1263 Collaboratory 2014).

(FDA, United States. Guidance Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products)

Data Consistency:

Relevant uniformity in data across clinical sites, facilities, departments, units within a facility, providers, or other assessors.

(FDA, United States. Guidance Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products)

Data Curation:

The curation of the source data for the purpose of statistical analysis of specific clinical research questions. Data curation includes, but is not limited to, the following aspects: data extraction (including multiple data sources), data security processing (de-identification or anonymization, and protection from data corruption, leaking, theft, tampering, or unauthorized access), data

cleaning (edit check and outliers processing, data completeness processing), data conversion (common data models, normalization, natural language processing, medical coding, derived variable calculation), data quality control, data transmission and storage.

(NMPA, China. Guideline on Using Real-World Data to Generate Real-World Evidence (Trial Version) English Translation)

Data Holder:

A legal person, including public sector bodies and international organizations, or a natural person who is not a data subject with respect to the specific data in question, which, in accordance with applicable law, has the right to grant access to or to share certain personal data or non-personal data

(Article 2(8) of the REGULATION (EU) 2022/868 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act))

Data Provenance:

An audit trail that “accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.”

(FDA, United States. Guidance Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products)

Data Relevance:

Data relevance includes the availability of key data elements (exposure, outcomes, covariates) and sufficient numbers of representative patients for the study.

(FDA, United States. Guidance Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision Making for Drug and Biological Products)

Data Reliability:

Data reliability includes data accuracy, completeness, provenance, and traceability.

(FDA, United States. Guidance Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products)

Data Traceability:

Permits an understanding of the relationships between the analysis results (tables, listings, and figures in the study report), analysis datasets, tabulation datasets, and source data.

(FDA, United States. technical specifications document Study Data Technical Conformance Guide (October 2019))

Digital Health Technology (DHT):

A system that uses computing platforms, connectivity, software, and/or sensors for health care and related uses. These technologies span a wide range of uses, from applications in general wellness to applications as a medical device. They include technologies intended for use as a medical product, in a medical product, or as an adjunct to other medical products (devices, drugs, and biologics). They may also be used to develop or study medical products.

(FDA, United States. Digital Health Technologies for Remote Data Acquisition in Clinical

Investigations Guidance for Industry, Investigators, and Other Stakeholders)

Effect Modification:

Effect modification occurs when the effect of a single exposure on an outcome depends on the values of another variable, i.e., the effect modifier, which does not necessarily need to be involved in the causal pathway. Interaction occurs when there is interest in the causal effect of two exposures on an outcome and how the effect of either exposure depends upon the value of the other exposure.

(ENCePP)

Electronic Health Record Data:

A collection of an individual patient records contained within an EHR system. A typical individual EHR may include a patient's medical history, diagnoses, treatment plans, immunization dates, allergies, radiology images, pharmacy records, and laboratory and test results.

(FDA, United States. Data Standards for Drug and Biological Product Submissions Containing Real-World Data)

Estimand:

A precise description of the treatment effect reflecting the clinical question posed by the trial objective. It summarizes, at a population level, what the outcomes would be in the same patients under different treatment conditions being compared.

(ICH E9-R1 - Addendum: Statistical Principles for Clinical Trials, Glossary).

Exposure:

An exposure is the medicinal product or regimen of interest being evaluated in the proposed study (ICH M14 Expert Working Group).

Federated Data Network:

A series of decentralized, interconnected nodes, which allows data to be queried or otherwise analyzed by other nodes in the network without the data leaving the node it is located at. Examples of FDNs include DARWIN EU, Sentinel, CNODES, OHDSI, and MID-NET.

(Hallock H, Marshall SE, 't Hoen PAC, Nygård JF, Hoorne B, Fox C, Alagaratnam S. Federated Networks for Distributed Analysis of Health Data. *Front Public Health*. 2021;9:712569.)

Medical Claims Data:

A compilation of information on medical claims submitted to insurance companies for reimbursement of medical expenses for treatments and other interventions. Medical claims data use standardized medical codes, such as the World Health Organization's International Classification of Diseases Coding (ICD-CM) diagnosis codes, to identify diagnoses and treatments.

(FDA, United States. Data Standards for Drug and Biological Product Submissions Containing Real-World Data)

Medicine:

Any substance or combination of substances intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease.

(Section 201(g) of the Federal Food Drug and Cosmetic Act (FD&C Act).)

Operational Definition:

The data-specific operation or procedure a researcher followed to measure constructs in a particular study.

(FDA, United States. Guidance Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products (Draft))

Patient Experience Data

Data that are collected by any persons and are intended to provide information about patients' experiences with a disease or condition. Patient experience data can be interpreted as information that captures patients' experiences, perspectives, needs, and priorities related to (but not limited to): 1) the symptoms of their condition and its natural history; 2) the impact of the conditions on their functioning and quality of life; 3) their experience with treatments; 4) input on which outcomes are important to them; 5) patient preferences for outcomes and treatments; and 6) the relative importance of any issue as defined by patients.

(Title III, section 3001 of the 21st Century Cures Act, as amended by section 605 of the FDA Reauthorization Act of 2017 [FDARA])

Phenotype / Phenotype Algorithm:

Observable and measurable information that is relevant to health or healthcare such as a disease (e.g., type 2 diabetes), a blood pressure measurement, a blood sugar value or an antibiotic prescription. It can be used to define any patient characteristics, from exposure to outcome. The translation of the case definition into an executable algorithm that involves querying clinical data elements from the EHRs is the Phenotyping algorithm. These algorithms identify and extract data from health records using clinical codes (for example ICD-10 or SNOMED). They can also be referred to as “electronic phenotype” or “computable phenotype”.

(www.ohdsi.github.io (The Book of OHDSI))

Plausibility:

The believability or truthfulness of data values.

(FDA, United States. Guidance Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products citing Kahn et al. 2016).

Primary Data Collection:

Data collected specifically for the present study.

(Adapted from ICH E8)

Quality Assurance (QA):

All those planned and systematic actions that are established to ensure that the *study* is performed and the data are generated, documented (recorded), and reported to an appropriate quality standard and applicable regulatory requirements.

(Adapted from E6(R2) Good Clinical Practice (GCP) - *Step 4* (final); 9 November 2016 – Glossary)

Quality Control (QC):

The operational techniques and activities undertaken within the quality assurance system to verify that the requirements for quality of the study-related activities have been fulfilled.

(Adapted from E6(R2) Good Clinical Practice (GCP) -- *Step 4* (final); 9 November 2016 – Glossary)

Quantitative Bias Analysis:

Quantitative bias analysis is an overarching term applied to methods that estimate quantitatively the direction, magnitude, and uncertainty associated with systematic errors that influence measures of associations.

(Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative Bias Analysis in Regulatory Settings. *Am J Public Health*. 2016;106(7):1227-30.)

Real-World Data (RWD):

Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.

Examples of RWD include data derived from electronic health records (EHRs); medical claims and billing data; data from product and disease registries; patient-generated data, including from mobile devices and wearables; and data gathered from other sources that can inform on health status (e.g., genetic and other biomolecular phenotyping data collected in specific health systems).

(Adapted from FDA, United States. Guidance Real-World Data: Assessing Registries To Support Regulatory Decision-Making for Drug and Biological Products DECEMBER 2023 and FDA, United States. Draft Guidance Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision Making for Drug and Biological Products)

Real-World Evidence

The clinical evidence about the usage and potential benefits or risks of a medicinal product derived from analysis of RWD.

(FDA, United States. Guidance Real-World Data: Assessing Registries To Support Regulatory Decision-Making for Drug and Biological Products DECEMBER 2023)

Registry:

A registry is an organized system that collects prespecified uniform data from a population defined by a specific disease, condition, or exposure.

(Adapted from: FDA Real-World Data: Assessing Registries To Support Regulatory Decision-Making for Drug and Biological Products DECEMBER 2023 and EMA Guideline on registry-based studies 24 September 2020)

Secondary Use of Data:

Use of existing data for a different purpose than the one for which they were originally collected.

(EMA Guideline on registry-based studies)

Standard of Care:

Treatment that is accepted by medical experts as a proper treatment for a certain type of disease or condition and that is widely used by healthcare professionals. Also called best practice, standard medical care, or standard therapy.

(National Cancer Institute Dictionary)

Statistical Analysis Plan:

A statistical analysis plan is a document that contains a more technical and detailed elaboration of the principal features of the analysis described in the protocol, and includes detailed procedures for executing the statistical analysis of the primary and secondary variables and other data.

(E9 Statistical Principles for Clinical Trials -- Step 4 (final); 5 February 1998 – Glossary)

Target Trial:

A hypothetical randomized trial that would answer the question of interest if it were feasible.

(Adapted from: National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Care Services; Committee on Developing a Protocol to Evaluate the Concomitant Prescribing of Opioids and Benzodiazepine Medications and Veteran Deaths and Suicides. An Approach to Evaluate the Effects of Concomitant Prescribing of Opioids and Benzodiazepines on Veteran Deaths and Suicides. Washington (DC): National Academies Press (US); 2019 Sep 24. 2, Specifying the Target Trial.)

1072 **13 Regulatory Guidelines Referenced**

- 1073 European Medicines Agency. Clinical Trials Regulation (EU) NO 536/2014; 2014 Apr.
- 1074 European Medicines Agency. Guideline on Good Pharmacovigilance Practices (GVP) –
1075 Module III. Revision 1; 2014 Sep.
- 1076 European Medicines Agency. Guideline on Good Pharmacovigilance Practices (GVP) –
1077 Module IX. Revision 1; 2017 Oct.
- 1078 European Medicines Agency. Guideline on Registry Based Studies. Amsterdam (NL); 2021
1079 Oct.
- 1080 European Medicines Agency. Guideline on the Content, Management and Archiving of the
1081 Clinical Trial Master File (Paper and/or Electronic). London (GB); 2018 Dec.
- 1082 European Medicines Agency. ICH E2D Post-Approval Safety Data Management. London
1083 (GB); 2004 May.
- 1084 European Medicines Agency. ICH E8 (R1) General Considerations for Clinical Studies.
1085 Revision 1. Amsterdam (NL); 2021 Oct.
- 1086 European Medicines Agency. ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis
1087 in Clinical Trials to the guideline on statistical principles for clinical trials. Amsterdam (NL);
1088 2020 Feb.
- 1089 European Medicines Agency. ICH Guideline for Good Clinical Practice E6(R2). London
1090 (GB); 2016 Dec.
- 1091 European Medicines Agency. ICH Topic E9 Statistical Principles for Clinical Trials. London
1092 (GB); 1998 Sep.
- 1093 European Medicines Agency. Patient experience data in EU medicines development and
1094 regulatory decision-making. Outcome of the workshop on 21st September 2022. 17 October
1095 2022, EMA/786952/2022.
- 1096 European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP).
1097 ENCePP Checklist for Study Protocols. Revision 4. EMA; 2018 Oct.
- 1098 European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP).
1099 Guide on Methodological Standards in Pharmacoepidemiology. Revision 11. Amsterdam
1100 (NL); EMA; 2023 July.
- 1101 FDA, United States. Draft Guidance for Industry: Real-World Data: Assessing Electronic
1102 Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug
1103 and Biological Products; 2021 Sep.
- 1104 FDA, United States. United States. Final Guidance: Best Practices for Conducting and
1105 Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data; 2013
1106 May.
- 1107 FDA, United States. Framework for FDA's Real-World Evidence Program. Silver Spring
1108 (US); 2018 Dec.

ICH M14 Guideline

- 1109 FDA, United States. Guidance for Industry: Considerations for the Use of Real-World Data
1110 and Real World Evidence To Support Regulatory Decision-Making for Drug and Biological
1111 Products; 2023 Aug.
- 1112 FDA, United States. United States. United States Guidance for Industry: Data Standards for
1113 Drug and Biological Product Submissions Containing Real-World Data; 2023 Dec.
- 1114 FDA, United States. Guidance for Industry: Real-World Data: Assessing Registries to
1115 Support Regulatory Decision-Making for Drug and Biological Products; 2023 Dec.
- 1116 FDA, United States. Patient-Focused Drug Development Guidance Series for Enhancing the
1117 Incorporation of the Patient’s Voice in Medical Product Development and Regulatory
1118 Decision Making; 2023 July.
- 1119 Health Canada, Canada. Elements of Real-World Data/Evidence Quality Throughout the
1120 Prescription Drug Product Life Cycle; 2019 Mar.
- 1121 ICH. ICH Reflection Paper Proposed ICH Guideline Work to Advance Patient Focused Drug
1122 Development; 2021 Jun.
- 1123 NMPA, China. Guidance for Real-World Data Used to Generate Real-World Evidence; 2021
1124 Apr.
- 1125 NMPA, China. Guidance on the Use of Real-World Evidence to Support Drug Development
1126 and Regulatory Decisions; 2020 Jan.
- 1127 NMPA, China. Guidance on Using of Real-World Study to Support Pediatric Drug
1128 Development and Regulatory Evaluation; 2020 Aug.
- 1129 MHLW/PMDA, Japan. Basic principles on the utilization of health information databases for
1130 Post-Marketing Surveillance of Medical Products; 2017 Jun.
- 1131 MHLW/PMDA, Japan. Guidelines for the Conduct of Pharmacoepidemiological Studies in
1132 Drug Safety Assessment with Medical Information Databases; 2014 Mar.
- 1133 World Health Organization. A Practical Handbook on the Pharmacovigilance of Medicines
1134 Used in the Treatment of Tuberculosis: Enhancing the Safety of the TB patient. Geneva (CH):
1135 WHO Press; 2012.
- 1136 www.fda.gov (Patient-Focused Drug Development Glossary 2018.)
- 1137 www.fda.gov FDA Patient-Focused Drug Development Guidance Series for Enhancing the
1138 Incorporation of the Patient’s Voice in Medical Product Development and Regulatory
1139 Decision Making 2024.
- 1140 www.imi-prefer.eu (How to do a patient preference study.)
- 1141 www.reganudall.org (Real World Data: Assessing Electronic Health Records and Medical
1142 Claims Data to Support Regulatory Decision Making for Drug and Biological Products
1143 Guidance for Industry 2021 Nov.)
- 1144 **14 Non-regulatory Guidelines Referenced**
- 1145 International Society for Pharmacoepidemiology. Guidelines for Good

ICH M14 Guideline

- 1146 Pharmacoepidemiology Practices (GPP). Washington, DC (US): ISPE; 2015 Jun.
- 1147 The REporting of studies Conducted using Observational Routinely-collected health Data
1148 (RECORD) Statement.
- 1149 The REporting of studies Conducted using Observational Routinely-collected health Data
1150 Statement for Pharmacoepidemiology (RECORD-PE).
- 1151 “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly work in
1152 Medical Journals,” established by the International Committee of Medical Journal Editors
1153 (ICMJE).

1154 **15 References**

- 1155 1. Wang SV, Pottegard A, Crown W, Arlett P, Ashcroft DM, Benchimol EI, et al.
 1156 HARmonized Protocol Template to Enhance Reproducibility of Hypothesis
 1157 Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices
 1158 Report of a Joint ISPE/ISPOR Task Force. *Pharmacoepidemiol Drug Saf.*
 1159 2023;32(1):44-55.
- 1160 2. Gatto NM, Campbell UB, Rubinstein E, Jaksa A, Mattox P, Mo J, et al. The
 1161 Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment
 1162 Framework. *Clin Pharmacol Ther.* 2022;111(1):122-34.
- 1163 3. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, Dal Pan G,
 1164 Goettsch W, Murk W, Wang SV. Graphical Depiction of Longitudinal Study Designs
 1165 in Health Care Databases. *Ann Intern Med.* 2019 Mar 19;170(6):398-406).
- 1166 4. Hernan MA, Robins JM. Using Big Data to Emulate a Target Trial When a
 1167 Randomized Trial Is Not Available. *Am J Epidemiol.* 2016;183(8):758-64.
- 1168 5. Wang SV, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin JA, Bartels DB, Kahler KH,
 1169 Bessette LG, Schneeweiss S. STaRT-RWE: structured template for planning and
 1170 reporting on the implementation of real world evidence studies. *BMJ.* 2021 Jan
 1171 12;372:m4856. doi: 10.1136/bmj.m4856. PMID: 33436424; PMCID: PMC8489282.