M14 General Principles on Planning, Designing, Analyzing, and Reporting of Non-interventional Studies That Utilize Real-World Data for Safety Assessment of Medicines

Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

March 2026
ICH-Multidisciplinary

M14 General Principles on Planning, Designing, Analyzing, and Reporting of Non-interventional Studies That Utilize Real-World Data for Safety Assessment of Medicines
Guidance for Industry

Additional copies are available from:

Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
Phone: 855-543-3784 or 301-796-3400
Email: druginfo@fda.hhs.gov
https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs

and/or

Office of Communication, Outreach and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
Phone: 800-835-4709 or 240-402-8010
Email: industry.biologics@fda.hhs.gov
https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

March 2026
ICH-Multidisciplinary

Contains Nonbinding Recommendations

FOREWORD

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) has the mission of achieving greater regulatory harmonization worldwide to ensure that safe, effective, and high-quality medicines are developed, registered, and maintained in the most resource-efficient manner. By harmonizing the regulatory expectations in regions around the world, ICH guidances have substantially reduced duplicative clinical studies, prevented unnecessary animal studies, standardized safety reporting and marketing application submissions, and contributed to many other improvements in the quality of global drug development and manufacturing and the products available to patients.

ICH is a consensus-driven process that involves technical experts from regulatory authorities and industry parties in detailed technical and science-based harmonization work that results in the development of ICH guidances. The commitment to consistent adoption of these consensus-based guidances by regulators around the globe is critical to realizing the benefits of safe, effective, and high-quality medicines for patients as well as for industry. As a Founding Regulatory Member of ICH, the Food and Drug Administration (FDA) plays a major role in the development of each of the ICH guidances, which FDA then adopts and issues as guidance to industry.

TABLE OF CONTENTS

Contains Nonbinding Recommendations

M14 General Principles on Planning, Designing, Analyzing, and Reporting of Non-interventional Studies That Utilize Real-World Data for Safety Assessment of Medicines Guidance for Industry[1]

This guidance represents the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA office responsible for this guidance as listed on the title page.

I. INTRODUCTION (1)

A. Objectives (1.1)

The purpose of this guidance is to recommend international standards for, and promote harmonization of, the general principles on planning, designing, analyzing, and reporting of non-interventional studies that utilize ***fit-for-use*** (frequently referred to as fit-for-purpose) data for safety assessment of ***medicines*** (drugs, vaccines, and other biological products). The Glossary defines many of the terms for the purpose of this of this guidance. Words or phrases found in the Glossary appear in bold italics at first mention.

Broadly, pharmacoepidemiology is a scientific discipline that uses epidemiological methods to evaluate the use, benefits and risks of medicines, medical technologies, and other interventions in human populations (Ref. 1). This document outlines recommendations and high-level best practices for the conduct of non-interventional pharmacoepidemiologic safety studies, hereafter referred to as non-interventional studies (as further defined in Section I.C., Scope (1.3), to streamline the development and regulatory assessment of study protocols and reports. These recommendations and practices also seek to minimize the conduct of multiple studies on the same safety concern for submission to multiple regulators, and to improve the ability of the study protocol to be accepted across regulatory authorities, and support decision-making in response to study results. Terms that appear in ***bold italic*** type upon first use are defined in the Glossary.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only

[1] This guidance was developed within the Expert Working Group (*Multidisciplinary*) of the International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) and has been subject to consultation by the regulatory parties, in accordance with the ICH process. This document has been endorsed by the ICH Assembly at *Step 4* of the ICH process, September 2025. At *Step 4* of the process, the final draft is recommended for adoption to the regulatory bodies of the ICH regions.

as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word should in Agency guidances means that something is suggested or recommended, but not required.

B. Background (1.2)

Non-interventional studies have long been a source of evidence to support the evaluation of the post-marketing safety of approved medicines.

Safety concerns can arise from a variety of data sources; these sources may include all clinical and scientific information concerning the use of medicines and the outcome of this use, such as product quality data, non-clinical studies, clinical trials, pharmacovigilance data, and non-interventional studies. Non-interventional studies are a key component in the detection, characterization, and evaluation of safety concerns, and may be descriptive or inferential.

Generation of robust evidence to be used for regulatory purposes depends on the ***reliability*** and ***relevance*** of the data and the application of sound methods to analyze such data. The use of non-interventional studies for regulatory decision-making has increased globally, and multiple guidances and best practice documents have been developed by regulators and professional societies. Many countries and regions have published guidances related to general principles of planning and designing such studies, mainly for the purpose of safety assessment of a medicine. In addition, frameworks for study design, conduct, and reporting are being developed by non-governmental groups, and may be used to help guide development of a study. These include the European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology, The Sentinel Innovation Center with the PRINCIPLED framework,(2) and ISPE/ISPOR's HARmonized Protocol Template to Enhance Reproducibility (HARPER) Initiative, which provide additional detail that is beyond the scope of this guidance (Refs. 3-6).

C. Scope (1.3)

Although there are slight differences between regions with regard to what constitutes ***real-world data*** (RWD), this guidance provides recommendations for the generation of ***real-world evidence*** (RWE) that is submitted to regulators for the purpose of evaluating post-marketing safety of medicines. At times, RWD sources alone may be insufficient to answer the research question of interest, and a study will require additional data for the purposes of the study. Because ***primary data collection*** may be relevant to non-interventional studies using RWD, this guidance also includes considerations for primary data collection (see the guidance for industry *ICH E8 General Considerations for Clinical Studies* for additional information on this topic (Ref. 7)).[2]

Collecting ***patient experience data*** may be a valuable component for post-marketing safety studies to inform on aspects such as notable events, perspectives, needs, and priorities. While a detailed guidance on this is beyond the scope of this guidance, several regulatory guidances have

[2] We update guidances periodically. For the most recent version of a guidance, check the FDA guidance web page at https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

been developed. When studies include patient experience data, the researcher(s) may consult relevant published recommendations for additional information.

It is beyond the scope of this document to provide guidance on whether a clinical trial or a non-interventional study is the most appropriate approach, nor is it intended as a comprehensive source of knowledge for non-interventional safety studies. Rather, the intent is to harmonize regulatory guidance documents for the design, planning, analysis, and reporting of non-interventional studies, and to facilitate regulatory review. The researcher(s) can also consider, as relevant, best practice guidances from other sources to the extent not covered in regulatory guidance. Similarly, citations presented are not intended to be prescriptive nor exhaustive, but illustrative for key concepts.

The following are out of scope:

- Pharmacovigilance studies relying only on routine spontaneous reports obtained from national or regional data sources (e.g., pharmacovigilance systems at national level);

- Studies involving treatment assignment, including randomized controlled trials, pragmatic trials, single arm clinical trials with treatment assignment defined per protocol, and trials using external comparators;

- Studies primarily involving user-generated health data extracted from other sources (e.g., websites, blogs, social media, chat rooms); and

- Studies evaluating the effectiveness of risk minimization programs (e.g., Risk Evaluation and Mitigation Strategy or additional Risk Minimization Measures studies), unless the evaluation takes the form of a non-interventional study to evaluate the safety concern.

The use of pharmacogenomics, artificial intelligence (AI), and other technologies may be relevant to the use of RWD and generation of RWE. However, considering the evolving landscape in these areas, this guidance does not address those topics.

D. Studies Conducted for Purposes Other Than the Safety Assessment of Medicines (1.4)

The principles presented in this document provide recommendations that may be applicable to non-interventional studies conducted for purposes other than evaluation of the safety of medicines, such as utilization and effectiveness studies or for the evaluation of the safety of medical devices. The basic principles presented in this guidance may be relevant to these studies when RWD are included.

II. GENERAL PRINCIPLES (2)

The safety profile of a medicine reflects an evolving body of knowledge extending from preclinical investigations through the post-marketing lifecycle. Non-interventional safety studies complement other sources of information such as spontaneous adverse event reporting and clinical trials to provide a better picture of the benefit-risk profile of a medicine as used in clinical practice.

The guidance describes a stepwise process, although the various steps of study design and data source selection are iterative. The process starts with articulating the study rationale and research question in response to a safety concern; then following a principled approach to identify the study population, exposure, comparator(s), outcome, and covariates; identifying the minimum data requirements to inform data source selection and guide feasibility assessment; assessing the representativeness of the data source to the target population; and considering sources of potential ***bias*** and ***confounding***. After an appropriate data source and or data collection approach has been identified, the process involves further refining the design, which includes approaches to address internal and external validity. Section III., Conceptual Framework for Generating Adequate Evidence Using Real-World Data (3) of this guidance describes the integration of these activities. Researcher(s) are encouraged to discuss the attributes of a particular study with regulators beginning early in the planning process. This could take the form of submission of a protocol synopsis, key design elements, or initial feasibility assessments provided to the regulators. Throughout, the underlying rationale and justification for exposure, outcome, and covariate definitions, analysis, data management, study implementation, reporting and regulatory submission, and other key decisions should be documented and discussed with regulators as appropriate.

In this guidance we refer to "researcher(s)" as those responsible for designing and executing the study; this may be a regulatory agency, a sponsor or application holder, contract research organization, academic group, or others. Sponsors of marketing applications and marketing authorization holders (MAH) are ultimately responsible for all aspects of post-marketing safety studies submitted to regulators, including compliance with regulatory timelines, pathways, and other applicable guidances.

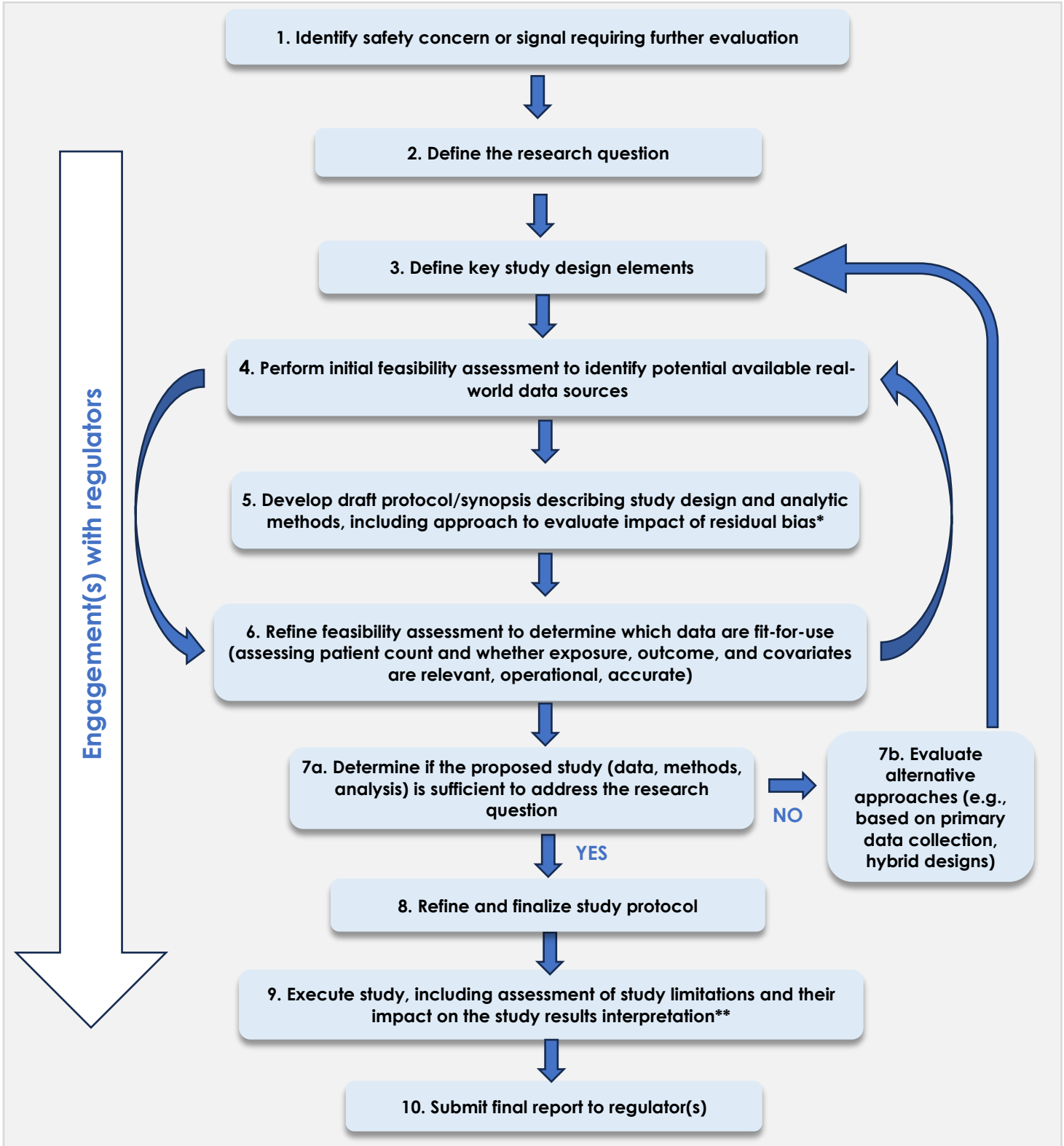
III. CONCEPTUAL FRAMEWORK FOR GENERATING ADEQUATE EVIDENCE USING REAL-WORLD DATA (3)

The strength of the study-generated evidence submitted in support of a regulatory decision depends on the research design and methodology in addition to the relevance and reliability of the underlying data. Figure 1 depicts a conceptual framework for designing and generating evidence that is adequate to address a research question in response to the emergence of a safety concern, including interactions with regulators during this process (Ref. 8). Although Figure 1 depicts a generally linear process (steps 1-10) for simplicity, it is in fact iterative, with some steps that may be done in parallel. Within this framework, once a safety concern has been identified and the research question established, then the study design and data source(s) most

appropriate for addressing those questions are determined (Ref. 8). The researcher(s) should design a study to address the research question, rather than designing a study to conform to a specific data source.

In order to determine if the evidence will adequately answer the research question, the framework depicted in Figure 1 provides for an integrated assessment of three components: (a) ***data relevance*** and ***data reliability***, (b) appropriateness of the study design and analytic methods, and (c) a robust qualitative/quantitative assessment of study limitations and their impact on the ultimate validity and reliability of the resulting evidence and the interpretation of findings. The integrated assessment of whether the evidence generated through the study will be adequate should be considered both during protocol development with a feasibility assessment (e.g., discussion of the impact of methodological concerns, consideration of possible sources of bias and their potential impact on study validity) and after study implementation with sensitivity analyses pre-specified in the protocol. ***Quantitative bias analysis*** is a set of methods that can be used to assess the sensitivity of study results to sources of systematic errors (e.g., misclassification, uncontrolled confounding, and selection bias), and one can further assess the impact of these biases on the direction and magnitude of effect estimates (Refs. 9-12). Such methods can be used in both study design and interpretation of results (Refs. 11-16). All three components considered jointly can enable a decision on whether the study, if executed according to the protocol, can generate adequate evidence to address the specific regulatory question, and in appropriate situations may be submitted to other regulators.

Figure 1. A conceptual framework for generating adequate evidence using fit-for-use real-world data to address regulatory questions on the safety of medicines



*Quantitative bias analysis methods can assist with study design and fit-for-use data evaluation to understand impact of bias due to misclassification or uncontrolled confounding on potential study sufficiency (Ref. 11.)
**Quantitative bias analysis can be used in the study analysis phase to evaluate the impact of residual or unmeasured confounding or misclassification on study interpretation (Ref. 12).

IV. INITIAL DESIGN AND FEASIBILITY (4)

A. Research Question (4.1)

The research question is a concise statement of the study purpose. If the study has causal (inferential) objectives, this statement may include prespecified null and alternative hypotheses. The research question may be formulated by use of the population, intervention (the exposure in the case of non-interventional studies), comparator, outcome, timing, and setting (PICOTS) template or similar structured templates, and should consider the setting in which the medicine is used (Refs. 17-18) When the research question is formulated, it should be explicit whether the study has descriptive or causal (inferential) objectives. In situations where an exposure-only study is conducted because a comparator cannot be identified (e.g., first-in-class or rare disease), the supporting rationale should be described. When applicable and possible, and to prevent the duplication of potential bias or confounding encountered in prior research, the specific question should be formulated after a critical appraisal of available information, including published literature to identify and understand any knowledge gaps, strengths and weaknesses of prior studies, the expected magnitude of the risk, and important confounding factors. When defining the research question, the researcher(s) should provide a clear rationale on how it will be addressed by the study objectives. Careful formulation of the research question will inform the feasibility assessment, and results of the assessment will provide information about feasible study designs and the type of data available to further refine the research question and subsequent protocol development. The researcher(s) may also consider a principled approach when formulating their research question. When the study has a causal (inferential) objective, an approach such as the target trial emulation may be used (Ref. 3).

B. Feasibility Assessment(s) (4.2)

A feasibility assessment is a systematic process to identify fit-for-use data to address a specific research question and to obtain information on the statistical precision of a potential study without evaluating outcomes between the exposed or comparator group(s). When conducting a feasibility assessment, a key goal is to describe and compare the relevance of the data source(s) assessed for the research question. Additional detail on potential strengths and limitations of data sources is provided in Section V., Protocol Development (5).

Feasibility assessments should be structured in at least two phases:

- An initial scan to determine whether data are available, likely sufficient, and to narrow down data source options; and

- A subsequent, more comprehensive feasibility assessment of the candidate data sources.

After the research question and design elements are established, the researcher(s) should specify the minimum criteria required to address the key design elements specific to the research question. This task will require an understanding of RWD source characteristics, including RWD relevance and reliability (see Section V.C., Data Sources (5.3)), and the clinical context. Elements to consider after defining the study objectives are the following:

- Study population (i.e., inclusion/exclusion criteria, index date, follow-up time, geography, extent of use (actual or anticipated) of the medication in the study population);

- Exposure (i.e., newly initiated vs prevalent use, where the patient obtained the medication);

- Comparator groups (i.e., the need for a comparator, availability, type [e.g., concurrent, historical]);

- Outcomes (i.e., available data/algorithms for operationalization or validation [see Section V.F., Validation of Key Variables (5.6)], adequate follow-up time);

- Covariates (i.e., available data/algorithms for operationalization); and

- Sample size required for desired statistical precision.

These elements are context-dependent and may narrow the available data sources, e.g., for rare diseases or rare outcomes. Several publications outline further details regarding features to consider (Refs. 8, 19, 20).

In the early stages of designing a non-interventional study, expectations regarding access to patient level or analytic data sets should also be clarified, including data handling and documentation that pertains to data reliability and provenance. Documentation includes procedures for data adjudication and verification as well as the transformation of data elements in the data source since these activities are often undertaken by ***data holder(s)*** and or aggregators (Refs. 8, 19, 20).

Depending on the research question, it may be appropriate to specify other important criteria, such as the ability to collect additional information to complement records in the data sources, or link data sources to other types of data (e.g., vital records, cancer registries, vaccine registries). At this point in the initial scan portion of the feasibility assessment, it should be possible to identify data sources that are most likely to satisfy the criteria the researcher(s) has/have specified as important to answer the research question. In some cases, it is possible for the researcher(s) to complete the initial scan step by solely relying on available information, including published literature, data source descriptions, catalogues of metadata, and occasionally, simple descriptive counts available from the data source.

Once a sufficient number of available data sources have been identified as potential candidates for utilization in the study, an in-depth feasibility assessment should be conducted. In some instances, fit-for-use data will be identified during the initial feasibility scan, in which case the detailed step will apply to the data sources under consideration. In the detailed feasibility step, the researcher(s) can verify that the specific data needed for the key design criteria are available and that there is sufficient evidence of validity and completeness of the minimum design elements in the specific data source.

When selecting a data source, data recency, frequency of data refresh, and completeness of observation time should be considered. Other factors in data source selection may include the researcher's prior experience with the data, how data are generated, whether data can be linked and or patients contacted, and ease of data access. There may be a trade-off between the time needed to address these factors versus the urgency of obtaining study results.

Analyses that evaluate the potential impact of data missingness may be conducted to further evaluate the feasibility of conducting a study in a given data source. A variety of information sources are used to complete this step, and it may often be valuable to request specific information from the data holder or an expert data user (e.g., number of patients fulfilling the inclusion and exclusion criteria (in total and by exposure group), incidence rate of outcome (in total) to conduct sample size calculations, availability of covariates, and other queries of the data to verify the data source is fit-for-use). This collaboration with a data holder or expert data user, who understands and has experience with data collection, data management or data analysis of the specific data source(s) of interest, may inform the process of identifying fit-for-use data.

After the detailed evaluation is completed, the data sources are compared, and data source(s) can be selected for the study (see Section V.C., Data Sources (5.3) for further detail on data sources). Occasionally, at any of the steps, it will be apparent that a specific data source is not suitable to address the research question. In these circumstances, the researcher(s) may conduct a feasibility assessment for other data collection options (e.g., primary data collection, or a hybrid approach), either on a subset of the cohort under assessment or a different, but similar, population for augmentation purposes. Primary data collection may be the only option available for some research questions and populations (e.g., rare diseases). This assessment typically includes physician and site queries, including information about the patient population (including potential biases introduced via volunteers), to determine if enough participants can be enrolled (based on factors like market uptake) and followed for the appropriate time frame to yield meaningful answers to the research question.

Whenever studies utilizing primary data collection are proposed, the researcher(s) should collaborate with study sites, data holders, and regulators to consider the time to set up the study, which includes time to select sites, train site staff, undergo ethics approval, enroll and follow participants, produce results, and whether this timing is acceptable, understanding that there may be no alternative. In-depth guidance on primary data collection is beyond the scope of this guidance, but additional information is available in the relevant chapters in Pharmacoepidemiology and the ENCePP Guide (Refs. 6, 21). Likewise, when using RWD sources, the researcher(s) should consider the time to set up the study agreement, data holder governance approvals, and data permit applications, to ensure that data are available in a timely manner.

In addition, the specification of an appropriate comparator group (or time period, operationalized as appropriate based on study design) is a critical part of the study design and an important consideration in the feasibility assessments as well as for comparative/inferential studies. The impact that disease severity of the exposed group and the comparator group have on policies for medical or medication coverage should be considered, as should the availability of concurrent

comparator data. However, in some situations (such as rare disease population studies) a historical or former ***standard of care*** comparator may be considered. In determining an appropriate comparator therapy, the researcher(s) should consider applicable regional regulatory guidance, as well as current clinical guidances and clinical studies in the present therapeutic indication.

Feasibility assessments are used as context for design decisions in the protocol. It is recommended that the researcher(s) note that a feasibility assessment has been conducted and includes the justification for a data source as part of the study files. As warranted, the details to be included and the appropriate format for submission of feasibility assessment results can be discussed with regulators. The description should include the objective of the feasibility assessment, and the data sources evaluated when designing the study, including results from feasibility evaluations or exploratory analyses of those data sources. The researcher(s) should provide a justification for selecting and excluding potential data sources for a feasibility assessment based on a priori information (i.e., limited sample size and data quality concerns) or requirements by a regulatory authority. Further, it is good practice to document that when the feasibility assessment was conducted, outcomes within treatment or comparison groups were not compared.

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

V. PROTOCOL DEVELOPMENT (5)

The design, conduct and interpretation of non-interventional safety studies should involve an experienced, multidisciplinary study team with the appropriate expertise in the following as applicable:

- Epidemiology and biostatistics, including study design (e.g., target population selection criteria, exposure, outcome, covariate definitions, and follow-up);
- Disease manifestation, causal pathways, and current clinical practices;
- The data sources selected;
- ***Data curation*** and management;
- Disease area insurance/billing and coding practices;
- Primary data collection (including patient consent, site selection and management, as necessary);
- Data privacy and security concerns (including accessing and sharing of healthcare data); and

- Ethics and institutional review board requirements.

Important elements to be considered in the study protocol include the choice of the data source(s) for the study, the completeness of data capture, variability among data sources, the impact of changes over time in the data source, governance and conditions of access to the data source, and the healthcare system of the country or region covered by the data source(s), assessment of the target population, exposure, comparator(s), outcome and covariates, as well as approaches to address bias and confounding.

Version control for the protocol is essential and is covered further in Section X., Study Documentation and Record Retention (10). Detailed templates and checklists for protocol development are available in regulatory guidance and scientific publications.

A. Study Design (5.1)

Non-interventional safety studies usually aim to estimate the incidence or prevalence of an adverse outcome in a population of interest and to evaluate its causal association with exposure to a medicine.

Several study designs are commonly used in non-interventional safety studies, including cohort, case-control, and self-controlled studies. The selection of the most appropriate study design depends on multiple factors including the research question of interest and what is known about the postulated relationship between exposure to a medicine and the specific safety outcomes of interest (e.g., acute versus latent outcome, biologic pathway).

Identifying the appropriate comparator population is a critical element of study design and should represent the counterfactual experience for the exposure under investigation. This is most directly accomplished by the cohort design, but other study designs may be more appropriate. Examples of comparator groups may include users of other medicines, non-users, historical controls, or the patient themselves in self-controlled designs. Considerations for comparator selection may include the specific indication within a disease, contraindications, disease severity or comorbidity, and the treatment sequence. It is important to maximize and evaluate the comparability of the exposed and comparator populations to reduce issues related to confounding. Secular changes (new drug approvals and market uptake) occurring outside the study can pose substantial challenges and should be considered in the choice of comparator. The researcher(s) should discuss their rationale for selecting a particular study design in the study protocol and final report. They should also consider developing graphical representations (such as a study design diagram) to clarify the study design and assessment periods such as inclusion period, look back period, follow-up period, overall study period, and calendar time related elements such as cohort identification period. Visualization of design details helps to clarify and communicate the study design to a broad audience of decision makers (Ref. 22). The proposed study design should be discussed with the regulatory agencies early in the process to ensure that the proposed study may provide adequate evidence for regulatory decision-making.

Following completion of the feasibility assessments, the proposed study (design, data, methods, analysis) should be prespecified in the protocol.

B. Target and Study Population (5.2)

The target population refers to the specific group of individuals that the research aims to study (e.g., children aged 12-16 years with attention deficit hyperactivity disorder). The study population is intended to be representative of the target population for which conclusions are intended to be drawn. The study population is defined via inclusion and exclusion criteria (e.g., demographic factors, medical conditions, disease status, severity, biomarkers) based on the following elements, among others:

- Study periods, such as index dates and look back period (e.g., to identify new users);
- ***Conceptual definitions*** and ***operational definitions*** for key variables (see Section IV.B., Feasibility Assessment[s] (4.2)) used to select the study population and how they should be validated (see Section V.F., 5.6, Validation of Key Variables (5.6)); and
- The completeness and accuracy of the data elements to fulfil the inclusion and exclusion criteria (see Section V.C.1., Characteristics of Data Source Types (5.3.1)).

C. Data Sources (5.3)

Before using any data source in support of regulatory decision-making, the researcher(s) should consider whether the data are fit-for-use by assessing its relevance and reliability (see Section III., Conceptual Framework for Generating Adequate Evidence Using Real-World Data (3), and Section IV.B., Feasibility Assessment[s] (4.2)). Data privacy and security concerns raised when accessing healthcare data sources should be considered. For the purposes of this guidance, the term relevance includes the availability of key data elements (patient characteristics, covariates, exposures, outcomes) and a sufficient number of representative patients for the study (target population); the term data reliability includes concepts such as ***data accuracy***, ***data completeness***, ***data traceability***, and ***data provenance***.

The protocol should describe and discuss the data source(s) used and how it/they are fit-for-use to address the research question of interest and should refer to documentation of key data characteristics. In addition, the protocol should state any coding systems used for classification of the exposure and outcomes (e.g., Anatomical Therapeutic Chemical [ATC] Classification System, International Classification of Disease [ICD], International Classification of Primary Care [ICPC]), and any methods used for data linkage and algorithms implemented to identify population, exposures and outcomes, and relevant covariates). Data collection methods and procedures should be described.

Several data source characteristics should be considered in non-interventional studies, as they may affect the study design and the interpretation of the results. These include differences in coding systems used across data sources, standardization of data elements, and settings of care captured (e.g., primary, hospital, specialty, rehabilitation). Patients, providers, or healthcare

systems may have different drivers (e.g., medical, monetary, social, cultural, healthcare access) for data collection or participation, and billing practices for reimbursement, which may impact the characteristics of the underlying data and further inform study design and interpretation.

For studies that use data from multiple data sources or study sites (e.g., ***Federated Data Networks*** [FDNs](23), meta-analysis, data pooling, or data linkage), the researcher(s) should describe the rationale and procedures for how data from different sources can be obtained and integrated with acceptable quality, given the potential for heterogeneity in population characteristics, clinical practices, and coding across data sources. When utilizing multiple data sources, the researcher(s) should consider traceability and steps taken to harmonize data across institutions or data sources. It may be necessary to address duplicated patient records and associated data. Some FDNs have been specifically designed to support scientific evaluations and regulatory decision-making, allowing a growing number of studies to include data from these FDNs, often from different countries. It is essential to understand the strengths and limitations of the chosen data source(s).

1. Characteristics of Data Source Types (5.3.1)

Examples of data source types include data derived from ***Electronic Health Records*** (EHRs), ***administrative claims data***, patient registries, patient-generated data, and data gathered from other sources that can inform on health status, such as interviews, mail surveys, computerized or mobile-application questionnaires, and measurements through ***digital health technologies*** (DHTs; see Data Collected by Digital Health Technologies). Regardless of the data source(s) used, information on the context of the evidence generation should be obtained (e.g., geographic location, setting in which the data were generated, period during which the data were collected, and demographic information distribution of populations included in the data source). A high-level summary of commonly used data source types is provided in the sections below.

Electronic Health Record Data

Electronic Health Record data are captured by healthcare institutions and generally include data related to medical care, including diagnoses, prescriptions, and laboratory tests. Prescribing data is well-captured, but it does not guarantee that the patient actually received or used the medication. Dispensing information may be included if the medication is dispensed within the medical institution. The administration of medications to patients in clinical settings (e.g., in-hospital infusion) may be available. Furthermore, it depends on the implementation of EHR systems and the operation of the medical institution, but there are cases where prescriptions for in-hospital medication are not always captured in the EHR as structured data, instead being captured as unstructured data in a separate department system or clinical notes. Therefore, it is important to understand the details of the data that can be used in the EHR. Given that components and formats of data may differ among medical institutions, the lack of standardized data formats is often a major issue in a study when integrating data from multiple institutions.

Key clinical information is often embedded in unstructured data within EHRs, either as free text fields (such as healthcare practitioner notes) or as other non-standardized information in digitized documents (such as PDF-based radiology reports). Free text may be used to further characterize

exposure and outcome (e.g., review of patient profiles) in EHR-based data sources. To enhance the efficiency of data abstraction, a range of approaches, including both existing and emerging technologies (e.g., natural language processing, computer vision for images or laboratory results evaluation) are increasingly being used to convert unstructured data into a computable, structured data format. When using unstructured data, it is important to check the reliability of the data and the appropriateness of the data conversion method.

When making secondary use of EHR data from multiple medical institutions, any differences in components and format of these data, including codes used (such as disease names, drug names, procedures, and laboratory test items) should be harmonized and the approach documented in the protocol. EHR data typically capture the medical encounter with the healthcare provider but may not reflect the actual delivery of healthcare (e.g., medicines that are ordered but not dispensed or administered) and may require additional linkage (e.g., to pharmacy records). In addition, obtaining comprehensive history of medicine use, or medical care data on patients with certain types of privacy concerns (e.g., sexually transmitted infection, substance use disorders, mental health conditions) can be challenging. Nevertheless, the unavailability of these data can result in inaccurate or incomplete data.

Administrative Claims Data

Healthcare claims data sources are often large and capture healthcare services for all individuals covered by a health insurance program(s). Typically, once claims for all healthcare provided to individuals within a health insurance program are fully adjudicated (i.e., final payment decisions made by insurance companies or claims processors), they are aggregated into a data source that reflects a more complete view of services. Some data sources will contain a mix of open (in-process) and closed (paid/denied) claims; the researcher(s) should understand the dynamic nature of the data in these cases. Without linkage to other data sources, it is often not possible to obtain information about healthcare visits, results from laboratory testing, outcomes of offspring in pregnancy studies, vaccinations, severity of the disease, injuries from accidents, and care not covered by health insurance. These issues may be due to numerous factors, including health insurance coverage policies, switching between health insurance providers, and seeking medical care outside of the insurance system (e.g., self-paid/self-care treatments, procedures insured by worker's accident insurance, and motor vehicle liability insurance). In publicly funded healthcare systems, public administrative claims data are characteristically available across a wide range of publicly funded health encounters and services. It is feasible to track individuals who utilize these services across contacts and service delivery continuously over relatively long periods of time.

Claims data generally include diagnostic codes and procedure codes used for primarily administrative purpose such as billing and reimbursement. EHR data may be valuable for validation of claims-based operational definitions (refer to Section V.F., Validation of Key Variables (5.6)).

Registries

Registries are organized systems that collect prespecified uniform data from a population defined by a specific disease, condition, or exposure, which define the characteristics for registry entry (Refs. 24, 25). For example, data are collected from patients with a certain disease or characteristic, such as pregnancy, lactation, a birth defect, or a molecular or genomic feature. Exposure-based registries are generally systems by which the researcher(s) collect data on patients exposed to a specific medicine or class of medicines.

Data collected in an already established registry may be used for reasons other than originally intended. Secondary use of registry data warrants the same considerations and fit-for-use assessment relevant to data sources such as EHR and claims data. The suitability of the registry to answer the research question should be evaluated, e.g., taking into account the registry population, data elements collected, length of follow-up, frequency of assessments, calendar time, level of data quality, and governance (including aspects on data sharing and data access). Additional considerations may include the type of registry and the impact of methods involved in patient selection on the representativeness of the population relative to the target population (such as geographic factors, total number of patients in the registry, number eligible, number of new patients entering the registry per year and number lost per year with reasons for exit). If data necessary to answer the research question(s) are not routinely collected within established registries, linkage to external data sources or supplemental data collection through other means (e.g., primary data collection) should be explored, such as when there is a need for an adequate comparator population in a study using an exposure registry that only accumulates data on the exposure of interest; or when key measures of exposure or covariates such as duration, dose and route of therapy administration, are missing. In some cases, existing registries may not be sufficient and establishing a new registry may be necessary.

Data Collected by Digital Health Technologies

Digital health technologies are systems that use computing platforms, connectivity, software, and or sensors for healthcare 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 medical products (devices and medicines). DHTs may also be used to develop or study medical products. Data collected by these technologies should be subject to the same fit-for-use assessments as other data sources. There may be a need to specify DHTs (e.g., version, software, hardware, manufacturer), or to harmonize data across different types of DHTs.

2. Research Networks (5.3.2)

Data Standardization

Data standardization is relevant to multi-data source studies, including research networks (e.g., federated data networks, ***common data models*** [CDMs]). There are several challenges to

consider in standardizing study data derived from RWD sources. These challenges to standardization include but are not limited to:

- The type of information the sources contain (e.g., disease definition, diagnostic criteria, procedures, medications);

- The variety of RWD sources and the level of consistency in formats and coding languages, including differences in source data captured regionally and globally using different languages, standards, and terminologies; and

- Differences in culture, healthcare systems, such as differences in data recording, business processes and local healthcare practice patterns, data source structure, vocabularies, coding systems, variable definition, and de-identification methodologies used to protect patient data when shared.

In addition, coding systems for diagnoses, medicines, and laboratory data, among others, are updated regularly. Therefore, plans for mapping coding systems as they evolve/change should be addressed at the protocol stage, including the approach used to map codes (automated, manual) and its limitations. It may be necessary to perform verification of the data standardization. Moreover, care should be given when reusing a code list from another study, as code lists reflect the individual study objectives, methods, and the time in which they were created.

A free-text/unstructured component exists in some data sources, and can be used to define inclusion criteria, exclusion criteria, exposures, outcomes, follow-up, and covariates. Each free-text component may be transferred into a structured table which prompts users to specify what is measured, the timing of measurement, the care setting, type of codes that are used to define the measure as well as the sources for any algorithms used to derive study measures, e.g., defining exposures, outcomes, or covariates. The process for creating a structured variable from unstructured data should be provided in the study documentation.

Federated Data Networks

FDNs enable distributed analyses by combining data or results across multiple data sources or multiple types of data (claims, EHR). When choosing to use an FDN for a study, there are specific issues unique to these systems that should be considered, such as the FDN's transformation of data into CDMs, and the differences between systems from which the data arise. Governance of an FDN (centralized or decentralized) also needs to be taken into account, as it has an impact on the operational aspects of a study (e.g., study design, planning, data acquisition, data quality and data standardization). Instead of using data transformed into CDMs, data that are harmonized based on a common protocol and ***statistical analysis plan*** (SAP) can be used to provide a standard structure for sharing and analyzing data.

Under the FDN framework, different approaches can be applied for combining data or results from multiple data sources. A common characteristic of all approaches is the fact that data partners maintain physical and operational control over electronic data in their existing

environment and therefore the data extraction is always done locally. Differences, however, exist in the following areas: use of a common protocol; use of a CDM; and where and how the data analysis is conducted. When data from multiple data sources are combined, measures should be taken to ensure that data on the same treatment episode for the same patient is not duplicated.

The choice of data captured in a CDM is optimized for the types of data measures and detail needed for the intended use. Therefore, data in CDM-driven networks rarely contain all of the source information present at the individual data sources, and the data elements chosen for a given CDM network may not be sufficient for all research purposes or questions. Furthermore, consideration should be given to variables in a CDM and how they may differ in collection and interpretation among different institutions.

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

- Decreasing the time to conduct a study, either through pre-developed analyses, or by increasing the size of study populations as this shortens the time needed to obtain the desired sample size. Large sample sizes may facilitate research on rare events, rare diseases, and less common exposures;

- Multi-data source studies may provide additional knowledge on whether a safety issue exists in different populations or across countries and thereby may reveal causes of differential effects, inform the generalizability of results, the consistency of information and the impact of biases on estimates; and

- Heterogeneity of treatment options and utilization patterns across institutions, communities or countries may allow for a more complete understanding of the effect of individual medicines.

Data Linkage

Data linkage can be used to increase the breadth and depth of data on individual patients over time and may be utilized to allow access to other data sources to support validation efforts. Linkage of data sources such as cancer or mortality registries to claims or EHR data may result in a higher quality study by including data not in the original data source. It is important to have a comprehensive understanding of the data and to assess the accuracy and completeness of the linkage and the resulting linked data as data linkage may present unique challenges. In some circumstances, chart review or text entries in electronic format linked to coded entries can be useful for exposure, outcome, and covariate identification. In addition, patient privacy and confidentiality should be carefully protected according to the applicable legal and ethical standards when linking multiple data sources.

Conceptually, a data linkage may be undertaken within a data source (e.g., mother–infant linkage) or across data sources (e.g., vital records, biobank). If the study involves a data linkage, the protocol should describe each data source, the information that will be obtained, linkage methods (e.g., deterministic linkage, probabilistic linkage), how patient privacy will be

safeguarded (such as using linkage tokens, and or other strategies for reducing re-identification risks) and the accuracy and completeness of data linkages over time. If the study involves generating additional data (e.g., interviews, surveys, computerized or mobile-application questionnaires, measurements through DHTs), the protocol should describe the methods of data collection and linkage, the data elements used for linkage, and what will be done if imperfect linkage exists, or contradictory data are found across linked data sources.

3. Missing Data (5.3.3)

Understanding what prompted the recording of data is important to assess missing data in the data source of interest. There are two scenarios where data can be missing. The first scenario is the data are intended to be collected but were not collected. The second scenario is the data are not intended to be collected in the data source and therefore not available. A record in EHR systems or administrative claims data sources is generated only if there is an interaction of the patient with the healthcare system. Lack of information such as a laboratory result or prescription could be caused by different circumstances, such as (a) it might not have been ordered by the healthcare provider; (b) it may have been ordered but not conducted or dispensed; (c) it may have been conducted, but the result (test, dispensing) is not recorded; or (d) there is evidence of the healthcare interaction and the result was stored in the data source, but data were not in an accessible format, or lost in the transformation and curation process when the final study-specific dataset was generated. Approaches to handle missing data are described in further detail in Section VII., Analysis (7).

4. Appropriateness of Data Sources in Addressing Safety Questions of Interest (5.3.4)

Once an evaluation of all data sources has been conducted, the researcher(s) should demonstrate an understanding of why the selected data source(s) are appropriate to address the specific research question(s). During development of the protocol, as informed by the feasibility assessment(s), the researcher(s) should describe the following key aspects of the proposed data source(s) to support the demonstration of their relevance, the rationale for data source(s) selection and how they might affect the generalizability of the study results to the target population:

- How well the selected data source(s) capture study elements (e.g., whether a variable is captured, and if so, the degree of completeness) which may be components of a future ***phenotype*** or ***phenotype algorithm***;

- As appropriate, the ability to validate the outcome and other key study elements (e.g., exposures, key covariates, inclusion/exclusion criteria) including phenotype algorithm (see Section V.F., Validation of Key Variables (5.6));

- Estimated sample size and expected precision/power;

- Generalizability to the target population;

- The historical experience with use of the selected data source(s) for research purposes, including references for publications citing relevant previous use for non-interventional studies which may demonstrate fit-for-use characteristics or other elements to support use of the data source(s) for the proposed study;

- Timing related to data availability, data refresh, and adjudication (if applicable);

- The relevant healthcare system factors, such as medication tiering (e.g., first-line, second-line), formulary decisions, and patient coverage, can influence the degree to which patients on a given therapy in one healthcare system might differ in disease severity, or other characteristics, from patients on the same therapy in another healthcare system;

- The key patient characteristics which might act as potential confounders, including demographics, health conditions, lifestyle factors, risk factors for the outcome, health system (e.g., private or public/governmental healthcare);

- Heterogeneity across RWD sources; and

- Potential limitations and strengths of the data source(s) for addressing the research question.

D. Exposures, Outcomes, Covariates (5.4)

If the initial feasibility assessment has indicated that the exposures, outcomes, and covariates of interest are likely to be adequately captured in the potential data sources, then defining and operationalizing these elements can proceed. This process generally starts with defining study constructs (e.g., exposures, outcomes, covariates) in general or qualitative terms, to create a conceptual definition, which should reflect current medical and scientific thinking regarding the variable of interest, such as: (a) clinical criteria to define a condition for population selection or as an outcome of interest or a covariate; or (b) measurement of drug intake to define an exposure of interest. The conceptual definition should include a detailed description of the data elements that would characterize the exposure, outcome, or covariates.

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

- Whether it is possible to translate a conceptual definition of the exposures, outcomes, and covariates into one that can be operationalized in selected data source(s);

- Whether the operational definition adequately captures all elements of the conceptual definition; and

- Whether the operational definitions and the performance characteristics (e.g., sensitivity, specificity, positive and negative predictive values, and kappa statistic) are adequate in the chosen data source(s) based on the research question (see Section V.F., Validation of Key Variables (5.6)).

The conceptual definition of a clinical condition may be referred to as the phenotype. The protocol and or the SAP should include a detailed description of the operational definition, sometimes referred to as the phenotype algorithm (including the coding system and rationale) and the associated limitations (e.g., measurement error, proxies), the potential impact of misclassification, and how these limitations can be mitigated through the study design and analysis. For unstructured data, a detailed description, rationale for use, search criteria to identify outcomes/exposures/covariates, and the list of codes or concepts should be provided. Operational definitions developed for one data source or study population might perform differently in other sources or populations in terms of sensitivity and specificity due to data source-specific characteristics as well as variations in the disease epidemiology across populations and data sources. If utilizing or adapting a definition used or validated in other studies or data sources, applicability should be justified.

When identifying exposures and outcomes in a data source for a specific study, these data are usually coded. When selecting a data source, appropriateness of the coding system for defining the exposures and outcomes should be confirmed.

The following elements warrant consideration during protocol development:

- Data source/type and structure;

- Development of exposure, outcome and covariate definitions and the method used to identify them;

- Development and performance of the operational definitions, including time points, data types (structured, unstructured), variable types (categorical, continuous), transformation of variable types, code types, mapping of dictionary codes (e.g., ICD-10 to MedDRA) when applicable, and the mechanism of evaluation (selection of gold standard) and performance measures;

- Mapping of the available data elements against those required for the research question;

- Documentation of variable validity and appropriateness of applying previously used algorithms to the data source/population of interest; and

- Potential impact of misclassification on study validity and interpretation.

1. Exposure (5.4.1)

Conceptual Definition

An exposure is the medicine or class of medicines of interest (and dosing or regimen) being evaluated in the proposed study. The product or class of medicines of interest is referred to as the treatment and may be compared to no treatment, standard of care, another treatment (for which the recommendations below would also apply), or a combination of the above.

Exposure definitions can have differing levels of granularity, such as ever exposed versus never exposed, duration of exposure, user type (e.g., incident versus prevalent), exposure windows (e.g., current versus past exposure) also referred to as risk periods or risk windows, multiple exposures (e.g., use of more than one medicine or concomitant vaccinations), treatment switching, sequencing (e.g., first line or second line) or dosage (e.g., current dosage, cumulative dosage over time). Consideration should be given to both the requirements of the study design and the availability of data. The exposure definition should include information about the medicine dose, brand, dosage form, strength, duration, as well as the route, timing, and frequency of administration (as applicable). It may also be necessary to describe the manufacturer, the excipients used, and the drug delivery device as part of the product identification (e.g., for a drug, vaccine, or other biological product with the same active substance name made by a different manufacturer). This may require an understanding of the pharmacological or biological properties of the medicine, or members of the product class, as well as how the medicine is prescribed and used.

Operationalizing Exposure

By Medicine Type, Route, and Setting

The protocol should include a discussion of strengths and limitations of translating the conceptual definition to the operational definition. Considerations may include:

- Medicines that are prescribed are not necessarily dispensed;
- Medicines that are dispensed are not necessarily used or administered;
- Patient adherence and ability to provide an accurate account of intake;
- Medicines that are not captured in the data source such as samples, low-cost medicines, non-prescription medicines, and vaccinations offered in the workplace; and
- Coding systems used to identify exposures (e.g., National Drug Code [NDC], ***RxNorm***, Healthcare Common Procedure Coding System [HCPCS], ATC Classification System, ***Procedure Codes***).

Medicines may be administered across multiple settings. For instance, infusions could be administered in private clinics or on an outpatient basis (e.g., home care) in addition to in-hospital. Thus, setting and treatment patterns should be considered carefully in terms of potential requirements for data linkages to avoid exposure misclassification.

For vaccines, it is important to have granular information on brand, dose schedule, coadministration with other vaccines, and batch number or administration route and site. These data are not typically available and may require linkage to vaccination registries.

By Medicine Dose, Timing, and Duration of Exposure

The exposure (i.e., dose, dosage regimen) to the medicine of interest should be well-defined and measured. If this type of information is not available in the data source, the protocol or study report should discuss the specific assumptions made when estimating the dose and regimen of the exposures of interest. Consideration should be given to the timing of exposure (e.g., the relevant exposure window, relative to the onset of the outcome), and this may be especially difficult when "as needed" or non-prescription medicines are an exposure of interest. When defining the exposure period, it is necessary to decide whether the start date of exposure is the date of prescription, the date of dispensing, or the date of administration. Because patients may not refill their prescriptions exactly on time or, alternatively, may refill their prescriptions early, gaps or stockpiling in therapy may exist and may be reflected in the data. Allowable gaps between dispensing to construct exposure episodes and the gaps between exposure episodes should be considered when determining whether an exposure period is continuous. Conditions for the completion of an exposure period should also be considered and explicitly defined (e.g., no record of a new prescription in the following six months), noting limitations such as the potential of a drug being prescribed to a patient in another setting that may not be captured in the dataset used for the study.

2. Outcome (5.4.2)

Conceptual Definition

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

Clinical outcome definitions should contain the source of the medical concept (e.g., search strategy defined clearly using, for example, Preferred term [PT] of the Medical Dictionary for Regulatory Activities [MedDRA] or the standardized MedDRA Query [SMQ], ICD code), diagnostic criteria, measuring methods and their quality control (if any), measurement tools (e.g., the use of questionnaire scales), calculation methods, measurement time points, variable types,

transformation of variable types (e.g., from quantitative to qualitative variable, if applicable), and the adjudication approach. If a new tool is being used in the study, then the validation information should be included, as applicable.

Operationalizing Outcome

An operational definition can be implemented using the data available, with acceptable performance to meet the objectives of the study. The conceptual definition is operationalized using diagnosis and procedure codes (e.g., ICD, Read, MedDRA), laboratory tests (e.g., Logical Observation Identifiers Names and Codes [LOINC]) and values, unstructured data or available variables without any coding needs (e.g., physician's encounter notes, radiology, or pathology reports), or measurement tools such as questionnaire scales (e.g., numerical pain scale). Consideration for changes in coding or the underlying EHR systems over time is essential. If unstructured data are used, detailed description and rationale for the methods and tools utilized and validation of those methods should be provided.

Single appearances of a diagnosis code may indicate a rule out diagnosis or lack adequate specificity. Instead, consider whether a valid definition of outcome can be achieved by combining medicines, laboratory data, and medical procedures used for diagnosis or treatment (e.g., at least two diagnosis appearances within a specified timeframe, a thromboembolism diagnosis code plus treatment with anticoagulant, anaphylaxis code plus use of epinephrine), rather than operationalizing the outcome only based on a single appearance of a diagnosis code. If the outcome is complex to define, information on the specialty of the physician making the diagnosis might provide reassurance regarding the quality of the information used to determine the outcome. Mortality as an outcome may not be included in electronic healthcare data unless the patient was under medical care at time of death. Linkage to external vital statistics resources may be necessary, though these data are also subject to other limitations such as availability and reliability of cause of death. Careful documentation of mortality data quality and its implications should be included in the protocol.

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

When conducting a study using data from multiple data sources (databases, institutions, sites), the researcher(s) should define the outcome considering the data differences between sources, such as diagnosis coding, laboratory reference ranges and medication records. A complete understanding of the timing and relationship between these elements is essential. In addition, having an accurate date of diagnosis and a clearly defined exposure-outcome risk window is

particularly important in order to account for biologically plausible timeframes for when the outcome, especially a long latency event, might be expected to occur if indeed associated with the exposure. The proposed outcome definition should distinguish disease onset (e.g., early symptoms) from a confirmed diagnosis with appropriate justification for accurate capture of disease onset.

When outcomes that measure patient experience are included (e.g., quality of life, subjective severity of disease), the protocol should specify how the outcome measure is defined, constructed, and validated, and the procedures for data collection.

The reason for the data collection and the nature of the healthcare system that generated the data should also be described as they can impact the quality of the available information and the presence of potential biases.

3. Covariates (5.4.3)

Covariates are variables that are neither an exposure nor an outcome of interest, and they either characterize a population or are a potential confounder or effect modifier to account for in the study design or analysis (see Section V.E., Bias and Confounding (5.5)). Relevant covariates may be known (measured or unmeasured) or unknown and their potential impact can be assessed using sensitivity analysis (see Section VII., Analysis (7)).

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

- **Confounding:** Confounding occurs when the estimate of measure of association is distorted by the presence of another factor. For a variable to be a confounder, it must be associated with both the exposure and the outcome, without being in the causal pathway.

Conceptual Definition

Covariates may be used to characterize cohorts, evaluate effect modification, and adjust for confounders (e.g., propensity scores, stratification, or matching). Definitions of covariates needed in a study should be prespecified based upon clinical, biological, psychological, and functional concepts, as appropriate. The definition should include a detailed description of the data elements used to construct the covariate.

Operationalizing Covariates

Moving from a conceptual to operational definition proceeds as with exposure and outcome. Covariates are typically identified and assessed during the period before the start of the exposure of interest (baseline). Assessment of baseline covariates can be performed using different periods of time. The length of this look-back period is selected by considering factors such as

changes in coding or medical practice, expected frequency of medical encounters, relevance to the research question, data availability, and the impact on study power. When medical practice and coding systems change over time, attention is warranted if comparing results to historical data. Covariates may also be assessed during the observation period, either as static or time-varying covariates. Reliable assessment of covariates is therefore essential for the validity of results, including the timing of assessments for each of the covariates. A given data source may or may not be suitable for studying a research question depending on the availability of information on these covariates. When the covariate is not available in the chosen data source, the researcher(s) may consider whether proxies for the covariate are appropriate. A clear reasoning should be provided to show that it is appropriate to use the proxy instead of the missing covariate. The researcher(s) should provide the developed operational definition, including codes and settings of care, for all covariates in the protocol.

E. Bias and Confounding (5.5)

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

The proposed data source should be evaluated to determine whether it is adequate to capture information on important factors so that bias and confounding may be adequately controlled or assessed. As discussed in Section III., Conceptual Framework for Generating Adequate Evidence Using Real-World Data (3), a plan to use quantitative bias analysis may be useful when evaluating the direction and magnitude of biases to inform strategies for bias mitigation, and how the study biases may influence the interpretation of the study (see Section VII., Analysis (7)). Linkage with other data sources or additional data collection to expand the capture of important variables that are unmeasured or imperfectly measured in the original data source should be considered. Biases specific to primary data collection should be considered (e.g., recall, volunteer, or interviewer bias). Sources of bias and confounding should be considered, and decisions to address should be justified during the design stage with a plan to evaluate the influence of bias and confounding; these should be included in the protocol, analysis plan or final report. Different types of bias and confounding are briefly described in the following subsections. Though outside the scope of this guidance, regional resources are available to provide information on bias and confounding (e.g., the ENCePP Guide on Methodological Standards in Pharmacoepidemiology (Ref. 6)).

1. Selection Bias (5.5.1)

There are different types of selection bias such as referral bias, self-selection bias, prevalent user bias, and differential loss to follow-up. Different forms of selection bias may be addressed in either the design or analysis stages; however, the researcher(s) should aim to address it during the design stage where it should also be considered when selecting the appropriate data source.

A common type of selection bias is prevalent user bias, which can arise when prevalent users of a medicine are included in a non-interventional study, i.e., patients already taking a therapy for some time before study follow-up began. Prevalent users are "survivors" of the early period of pharmacotherapy that is not captured in the study. This can introduce selection bias if the risk varies with time. For example, patients who initiated a new medicine, experienced a safety event, and then discontinued the medicine may not be included in the study, thereby leading to a potential underestimation of the risk among the treated.

2. Information Bias (5.5.2)

Information bias arises when misclassification of binary or categorical variables or mismeasurement of continuous variables exists. Examples include recall bias, protopathic bias, and surveillance (detection) bias. The misclassification of key variables should be minimized to accurately estimate the effect of exposure on the outcome. Overall, the extent of variable validation (see Section V.F., Validation (5.6)) should be determined by the necessary level of certainty and the implication of potential misclassification on study inference. As discussed in Section III., Conceptual Framework for Generating Adequate Evidence Using Real-World Data (3), a quantitative bias analysis may be useful when evaluating the direction and magnitude of biases to inform strategies for bias mitigation, and how the study biases may influence the interpretation of the study (see Section VII., Analysis (7)).

3. Time-Related Bias (5.5.3)

One potential type of time related bias may result from immortal time, which refers to a period of cohort follow-up time during which, because of the exposure definition, an outcome of interest cannot occur. Immortal time bias occurs when the person-time at risk is treated differentially between the two exposure groups through either misclassification or exclusion.
When aiming to avoid the risk of immortal time bias or other time-related biases, such as immeasurable time bias, selection of an appropriate index date is essential. These risks may be mitigated by design frameworks (see Section IV.A., Research Question (4.1)), as this approach aligns assessment of eligibility and baseline information with start of follow-up.

4. Confounding (5.5.4)

The researcher(s) is/are typically unable to capture all potential confounders that are relevant to a research question, introducing the potential for unmeasured or residual confounding. In pharmacoepidemiology, commonly considered confounding factors include demographics, indication for treatment, disease severity, previous medication use, comedications, comorbidities,

prognostic characteristics, frailty, lifestyle factors, and others, depending on the study question. A number of approaches are available to address or evaluate unmeasured confounding, including high dimensional propensity scores, negative controls, and linkages to external data sources such as surveys that include data on confounders unmeasured in the study data source. The presence and impact of potential confounding factors should be considered in the study design phase. Directed acyclic graphs can be used to understand the relations between the variables and identify potential confounding and intermediate effects in a longitudinal study, and the impacts of these assessed using quantitative bias analysis, as discussed in the Analysis section (see Section VII., Analysis (7)) (Ref. 26).

F. Validation of Key Variables (5.6)

Validity is the extent to which a concept (variable) is accurately measured in a study by the operational definition. Validation of exposure, outcome, and key covariates is important for internal validity of non-interventional studies (Ref. 27). There are various approaches to validation, which may be data-source specific. These may include complete verification, partial verification, clinical expert review, review of patient claims, or profile history. Validation efforts should be commensurate with the level of evidence required, such as validating the outcome variable for all potential cases or non-cases or verifying the performance of an operational definition to identify cases and non-cases. For data sources routinely used in research, documented validation of key variables may have been done previously. Any extrapolation of a previous validation study should however consider the effect of any differences in population disease prevalence, inclusion and exclusion criteria, the distribution and analysis of risk factors, and subsequent changes to healthcare, procedures, and coding. Sponsors should have early interactions with regulators to discuss and agree upon a proposed validation approach, such as partial versus full, or adoption of definitions validated previously. The description of the validation approach should include the data source, population, time frame, performance, reference standard, and a discussion of the applicability of the proposed operational definitions considering the level of evidence required.

When validating an operational definition, prespecify the metrics to be reported (e.g., sensitivity, specificity, positive predictive value, negative predictive value, kappa statistic, prevalence-adjusted and bias-adjusted kappa, intraclass correlation coefficient), and describe how they will be measured. The trade-off between false-positive and false-negative cases should be considered when selecting an operational definition and identifying the proper validation approach to support internal validity. If several operational definitions are under consideration, the performance of each should be evaluated and the potential bias should be assessed using quantitative bias analysis at the design stage. This is distinct from common sensitivity analyses conducted at the analysis stage. Validation studies, when conducted, should be described in the study protocol or as a separate document, as appropriate.

VI. DATA MANAGEMENT AND CURATION (6)

Appropriate data management for a non-interventional study depends on various factors, including the source of the data and the planned use of the study results. A data management

and or data curation plan should be developed prior to study initiation. The ***quality assurance*** (QA) and ***quality control*** (QC) plans should be developed before an analysis is undertaken and the various factors influencing quality (e.g., data management by the data holder, quality defects of the data, inadequate data processing and analysis, or inadequate training) should be identified and addressed to preserve the integrity of the study. Detailed quality standards to be fulfilled should be in accordance with local or regional regulatory requirements.

To facilitate regulatory review, where submission of datasets is a regulatory requirement, a description of the context, content, structure of files, and steps used to create the files should be included. Datasets should be retained in accordance with regulatory requirements in the region(s) to which they will be submitted.

A. Data Management Plan (6.1)

Data quality assurance processes, policies, and procedures should account for potential risks to data quality, including errors in interpretation or coding; errors in data entry, transfer, or transformation accuracy; errors in programming logic; inadequate training; data completeness; and data consistency.

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

B. Quality Assurance and Quality Control (6.2)

1. Data Quality Management (6.2.1)

Although data quality considerations may vary depending on the study design and data source, fundamental determinants of data quality at each step in the evidence generation process, such as the accuracy and plausibility of source data (e.g., reasonable values of age or lab values), completeness of data during extraction, data quality management, and governance and documentation need to be addressed before finalizing the protocol. Depending on the data source, data in a non-interventional study may lack strict quality control over the process of recording, collection, and storage. This can lead to incomplete data, missing key variables, or inaccurate records. The presence of such quality defects will affect subsequent data curation, applicability, and traceability of data. QA/QC considerations are important to the reliability of the data for regulatory decision making as described below (see Section VI.B.2, Data Holders (6.2.2); and (Section VI.B.3, Researcher[s] (6.2.3)).

2. Data Holders (6.2.2)

QA and QC procedures used by the data holders, should address the following considerations:

- Reliability of data collection and management;

- Frequency and type of any data error corrections or changes in data adjudication policies implemented by the data holders during the relevant period of data collection;

- Supportive documentation (e.g., peer-reviewed publications) examining data quality and or validity of the data sources;

- Updates and changes in coding practices (e.g., International Classification of Diseases, ICD codes) across the study period;

- Changes in the availability of key data elements during the study time frame and their potential effect on the study (e.g., when access to a key data feed is no longer available); and

- The extent of missing data over time (i.e., the percentage of data not available for a particular variable of interest) and procedures (e.g., exclusion, imputation) employed to handle these issues.

3. Researcher(s) (6.2.3)

While the data holder maintains control of the data and is responsible for the underlying data quality, the researcher(s) are responsible for aligning QC and QA procedures with data holders to ensure transparency, understanding of data strengths/limitations, and meeting the standards of quality criteria required by the regulators for non-interventional studies. Further, the researcher(s) are responsible for the management and QA of all data cleaning, processing, and analytic datasets. To balance the need for sufficient QA with reasonable resource expenditure for a particular purpose, a risk-based approach to QA is recommended. Issues that are essential to determining the reliability and relevance of the data should be addressed in the protocol, and include QA/QC procedures for data accrual, curation, and transformation into the final study-specific dataset.

The researcher(s) should implement and maintain QA/QC systems with written procedures. This is to ensure that studies are conducted, and results are generated, documented, and reported in compliance with the protocol, regional laws, ethical considerations, and the applicable regulatory requirement(s). Documentation of these processes may include, but are not limited to electronic documentation (i.e., metadata-driven audit trails, QC procedures) of data additions, deletions, or alterations from the data source to the final study analytic dataset(s). The researcher(s) should also document changes to data and the potential impacts of these changes for conducting this specific study. Methods for QA/QC of analytic programming should be described in the study documentation.

VII. ANALYSIS (7)

The analytic strategy includes descriptive and inferential analyses to address the study objectives, while accounting for potential sources of bias and confounding. In addition, the strategy should also include an empirical evaluation of unmeasured, mismeasured, or unknown

confounding and other sources of bias. The statistical analysis should be prespecified, reflect the information gained from the feasibility assessments(s), and be developed to meet the study objectives. An overview of the statistical analysis approach should be provided in the protocol. The complete SAP should be provided as a standalone document, or as a detailed section of the protocol. It is recommended to discuss the approach chosen with regulatory agencies. The SAP should provide sufficient detail to allow replication of the analyses to help ensure confidence in the results.

In some studies, data-driven analyses may be performed; it is important to distinguish between those that are pre-specified and those that are post-hoc. Pre-specified analyses should be documented in the protocol and analysis plan and deviations from the plan documented in the final report. Post-hoc analyses are often conducted in response to observations in the data to help in the interpretation of results and should be described and justified in the final report and interpreted with care.

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

Specific attention to the data management and analytical strategy is needed when conducting studies using multiple data sources.

A. Statistical Analysis (7.1)

1. Analytical Approach (7.1.1)

The analysis should be directed towards calculating an unbiased estimate (e.g., risk or rate differences and risk or rate ratios). Analyses should align with the research question (see Section IV.D., Research Question (4.1)). The analysis section is where a description and justification for the chosen approaches for the statistical analyses should be described, including the assumptions and conditions.

The following aspects and elements should be considered for inclusion in the statistical analysis, if appropriate: descriptive analyses, subgroup analyses, methods of estimation and associated assumptions needed for analysis, estimate of the anticipated study size/power/statistical precision, plans to address confounding and bias (e.g., selection bias, information bias, time-related bias, time-varying exposures, time-dependent confounders, and impact on validity of results), evaluation of effect modification, assessment of population comparability, sensitivity analyses, type I error control (e.g., for sequential analysis and multiple comparisons), assessment of generalizability to the population of interest, and plans for handling missing data.

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

2. Missing Data (7.1.2)

The researcher(s) should develop the protocol and the SAP with an understanding of the reasons for the presence and absence of information in the underlying data; consider data linkage and or imputation to address missing data, and address the implications of the extent of missing data on study findings (see Section V.E., Bias and Confounding (5.5)). Descriptive analyses should be included to characterize missing data. Assumptions regarding the missing data (e.g., missing at random, where the probability of missingness is influenced by the values of other variables, or missing not at random, where the probability of missingness cannot be explained by the values of other variables) underlying the statistical analysis for study outcomes and important covariates should be supported. The extent and implications of missing data on study findings should be described.

3. Sensitivity Analyses (7.1.3)

When planning for sensitivity analyses, a rationale for each analysis should be provided with its strengths and limitations. Sensitivity analyses should be conducted to assess bias and confounding by examining the effect of varying potentially critical study assumptions, such as those relating to design, exposure definition, outcome definition, missing data, and limitations of the data source(s) selected, and analysis approach. The analyses can facilitate better interpretation of study results in light of the extent of uncertainty noted. Sensitivity analyses should be pre-specified in the protocol and or the SAP with deviations documented in the final report.

Quantitative bias analysis, as a form of sensitivity analysis, evaluates the impact of potential bias on the measure of association. The protocol should pre-specify the indices (e.g., sensitivity, specificity, positive [PPV] and negative predictive values [NPV]) that will be used for quantifying bias and describe how the selected indices will be measured when validating variables of interest. The precision of the bias-adjusted effect estimates should be quantified using confidence intervals. These analyses may facilitate interpretation of study results.

VIII. REPORTING AND SUBMISSION (8)

A. Reporting of Adverse Events, Adverse Drug Reactions, and Product Quality Complaints (8.1)

According to the ICH guidance for industry *E2D on Post Approval Safety Data Management* (ICH E2D), adverse events, adverse drug reactions, other observations and product quality complaints identified during the conduct of a study may require reporting to the regulatory authority. Reporting requirements may vary by party (e.g., MAH, other sponsor or applicant, investigator, or independent research group) and by region, due to differences in regulatory reporting requirements. The ICH E2D provides guidance for MAHs on reporting of individual case safety reports for adverse events and adverse drug reactions. For other reporting

requirements (and for parties outside the scope of ICH E2D), refer to applicable laws and regulations and, as appropriate, pharmacovigilance guidances.

Depending on the type of data collection (primary data collection or ***secondary use of data***), regulatory requirements for reporting vary by jurisdiction. The identification, processing and reporting of adverse events should follow local reporting requirements and be described in the protocol.

B. Formatting and Content of Study Documents for Submission to Regulators (8.2)

Sponsors should refer to available guidance on structure and contents of study documents and have early discussion with the regulatory agencies regarding the required documents and timetables for submission, as applicable. These documents may vary based on applicable regulatory requirements, and can include the feasibility assessment, protocol, SAP, and progress/interim and final reports. In the absence of specific regulatory guidance, sponsors may utilize or adapt frameworks developed by the scientific community as a guide for document development, including but not limited to ISPE/ISPOR's HARmonized Protocol Template to Enhance Reproducibility (Ref. 4),

IX. DISSEMINATION AND COMMUNICATION OF STUDY MATERIALS AND FINDINGS (9)

For transparency, to support scientific exchange, and to allow the conduct of reproducible research, even where not mandated by regulatory requirements, the researcher(s) is/are encouraged to make the protocol publicly available in appropriate public registers (e.g., ClinicalTrials.gov, the HMA-EMA Catalogue of real-world data studies, or other registers where available) after protocol finalization. Registration of study reports may also be needed according to local regulatory requirements. Further vehicles for dissemination and communication of study results may include non-regulatory submission in scientific fora, scientific publications, and patient or practitioner-focused communications.

Several guidelines exist that provide recommendations for reporting studies in the scientific literature. These include "The REporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) Statement," "The REporting of Studies Conducted Using Observational Routinely Collected Health Data Statement for Pharmacoepidemiology (RECORD-PE)," ENCePP Guide on Methodological Standards in Pharmacoepidemiology (Ref. 6), Heads of Medicines Agencies – European Medicines Agency (HMA-EMA) Catalogues of Real-World Data Sources and Studies (Ref. 28) and "Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly work in Medical Journals," established by the International Committee of Medical Journal Editors (ICMJE) (Ref. 29). It is recommended that the results be made publicly available and communicated in an appropriate manner to study participants (as applicable). Communications should include a factual summary of the overall study results in an objective, balanced and nonpromotional manner, including relevant safety information and any limitations of the study.

X. STUDY DOCUMENTATION AND RECORD RETENTION (10)

Key documents and records related to the planning, conduct and results of a study should be kept in compliance with applicable standards and jurisdictional requirements. Key principles of study documentation are similar to those in the guidance for industry *E6(R3) Good Clinical Practice* (GCP) (Ref. 30) (especially for primary data collection) and Guidelines for Good Pharmacoepidemiological Practices (GPP) (Ref. 5) (especially for secondary use of data):

- Key documents and records, should be recorded, handled, stored, and archived in a way that allows its accurate reporting, interpretation, verification, and that ensures confidentiality and patient privacy in compliance with applicable privacy laws;

- Systems should be in place to ensure completeness of the study documentation, to enable version control, to prevent accidental or premature loss, prevent unauthorized access, alteration, destruction, disclosure or dissemination; and ensure that an audit trail is maintained;

- Systems should be in place with procedures that ensure the quality of every aspect of the documentation of study development, conduct, and reporting;

- Study information should be readily available and directly accessible upon request by regulators (e.g., internal or regulatory inspection ready) with risk-based quality checks or review processes to ensure that the primary record system is being maintained up-to-date and that all key documents are appropriately filed; and

- All information should be retained at least for the duration of time required by applicable regulatory requirements.

XI. CONSIDERATIONS IN SPECIAL POPULATIONS (11)

Special populations are often not enrolled or under-represented in pre-approval clinical studies and include subjects who are pregnant or lactating, infants, children, adolescents/young adults, older adults, immunocompromised patients, and people with disabilities or rare diseases. Therefore, post-marketing non-interventional studies may provide valuable information supporting the benefit/risk assessment of medicines in these populations. Designing such studies often requires unique considerations (e.g., when defining the study population) in addition to the concepts applying to any non-interventional study described in this guidance. Examples of methodological challenges include low incidence of exposures and or rare outcomes; multiple comorbidities and polypharmacy for older adults; difficulty in identifying cases or disease characteristics (e.g., duration and severity) in immunocompromised patients; identification of pregnancies and estimation of gestational age; and complexity and variety of maternal, pregnancy, delivery, embryo-fetal, and neonatal and childhood outcomes. These challenges may require linkage to complementary data sources, such as birth registries, pregnancy registries, and

patient registries. The researcher(s) are encouraged to consult scientific literature for publications addressing design and bias issues related to these specific populations (Refs. 31-33).

XII. GLOSSARY (12)

Administrative Claims Data: Data that arise from a person's use of the healthcare system and reimbursement of healthcare providers for that care. These data can consist of information on claims submitted to insurance companies for reimbursement of expenses for treatments and other interventions. Claims data use standardized codes, such as the World Health Organization's International Classification of Diseases Coding (ICD-CM) diagnosis codes, to identify diagnoses and treatments.

Bias: A systematic deviation in results from the truth (Ref. 34).

Case Definition: The clinical, biological, psychological, and functional concepts of the condition, which reflect the medical and scientific understanding of the condition.

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

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

Confounding: Confounding results from the presence of an additional factor, known as a confounder or confounding factor, which 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 (Ref. 6).

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).

Data Completeness: The "presence of the necessary data" (National Institutes of Health 1263 Collaboratory 2014).

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

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.

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.

Data Provenance: The origin of a piece of data and how it was recorded, captured, and added to the data source.

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

Data Reliability: Data reliability includes data accuracy, completeness, and traceability.

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.

Digital Health Technology: A system that uses computing platforms, connectivity, software, and or sensors for healthcare and related uses.

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 (Ref. 6).

Electronic Health Record: An Electronic Health Record (EHR) is an electronic version of a patient's medical history that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that person's care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports.

Exposure: In the context of this guidance, exposure is the medicine or regimen of interest being evaluated in the proposed study.

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 (Ref. 23).

Fit-for-Use: A determination of the relevance and reliability of a proposed data source for a given study.

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

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

Patient Experience Data: Data that are collected by any person 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 their 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.

Phenotype: Observable and measurable information describing patient characteristics that are 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.

Phenotype Algorithm: The translation of the phenotype (or case definition) into an executable algorithm using clinical data elements from the electronic healthcare data. They can also be referred to as "electronic phenotype" or "computable phenotype."

Primary Data Collection: Data collected specifically for the present study; definition adapted from ICH E8.

Procedure Codes: Procedure codes are standardized alphanumeric codes used in healthcare to identify medical and surgical procedures performed on patients. Procedure codes could be used for healthcare record documentation, insurance transactions, and data analysis. Commonly used procedure coding systems are the Current Procedural Terminology (CPT), Healthcare Common Procedure Coding System (HCPCS), and International Classification of Diseases (ICD).

Quality Assurance: 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.

Quality Control: 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.

Quantitative Bias Analysis: Quantitative bias analysis is a set of methods that can be used to assess the sensitivity of study results to sources of systematic errors (e.g., misclassification, uncontrolled confounding, and selection bias), and one can further assess the impact of these biases on the direction and magnitude of effect estimates (Ref. 11).

Real-World Data: Data relating to patient health status and or the delivery of healthcare 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).

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

Relevance: See *Data Relevance*.

Reliability: See *Data Reliability*.

RxNorm: A standardized terminology for medications from the National Library of Medicine.

Safety Signal: Information that arises from one or multiple sources that suggests a new potentially causal association, or a new aspect of a known association, between a medicine and an adverse event or set of related adverse events that is judged to be of sufficient likelihood to justify further evaluation (Adapted from ICH E2C).

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

Standard of Care: As defined in the National Cancer Institute Dictionary, 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.

Statistical Analysis Plan: 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.

XIII. LIST OF ABBREVIATIONS (13)

AI: Artificial intelligence
ATC: Anatomical Therapeutic Chemical [classification system]
CDM: Common data model
DHT: Digital health technology
EHR: Electronic health record
EMA: European Medicines Agency
ENCePP: The European Network of Centers for Pharmacoepidemiology and Pharmacovigilance
FDN: Federated Data Network
GCP: Guideline for Good Clinical Practice
GPP: Guidelines for Good Pharmacoepidemiology Practices
HARPER: HARmonized Protocol Template to Enhance Reproducibility
HCPCS: Healthcare Common Procedure Coding System
HMA: Heads of Medicines Agencies
ICD: International Classification of Disease
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICMJE: International Committee of Medical Journal Editors
ICPC: International Classification of Primary Care
ISPE: International Society for Pharmaceutical Engineering
ISPOR: The Professional Society for Health Economics and Outcomes Research
LOINC: Logical Observation Identifiers Names and Codes
MAH: Marketing authorization holder
MedDRA: Medical Dictionary for Regulatory Activities
NDC: National Drug Code
NPV: Negative predictive value
PICOTS: The population, intervention, comparator, outcome, timing, and setting template
PPV: Positive predictive value
PT: Preferred term
QA: Quality assurance
QC: Quality control
RECORD: The REporting of studies Conducted using Observational Routinely collected health Data [statement]
RWD: Real-world data
RWE: Real-world evidence
SAP: Statistical analysis plan
SMQ: Standardized MedDRA query

XIV. REFERENCES (14)

1. International Society for Pharmacoepidemiology (ISPE), 2025, Overview, accessed June 2, 2025, https://www.pharmacoepi.org/about-ispe/overview/.

2. Desai RJ, Wang SV, Sreedhara SK, Zabotka L, Khosrow-Khavar F, Nelson JC, Shi X, Toh S, Wyss R, Patorno E, Dutcher S, Li J, Lee H, Ball R, Dal Pan G, Segal JB, Suissa S, Rothman KJ, Greenland S, Hernán MA, Heagerty PJ, and S Schneeweiss, 2024, Process Guide for Inferential Studies Using Healthcare Data From Routine Clinical Practice to Evaluate Causal Effects of Drugs (PRINCIPLED): Considerations From the FDA Sentinel Innovation Center, BMJ, epub ahead of print Feb. 12, 2024, doi: 10.1136/bmj-2023-076460.

3. Hernan MA and JM Robins, 2016, Using Big Data to Emulate a Target Trial When a Randomized Trial is Not Available, Am J Epidemiol, 183(8):758-764.

4. Wang SV, Pottegard A, Crown W, Arlett P, Ashcroft DM, Benchimol EI, Berger ML, Crane G, Goettsch W, Hua W, Kabadi S, Kern DM, Kurz X, Langan S, Nonaka T, Orsini L, Perez-Gutthann S, Pinheiro S, Pratt N, Schneeweiss S, Toussi M, and RJ Williams, 2023, HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force, Pharmacoepidemiol Drug Saf, 32(1):44-55.

5. International Society for Pharmacoepidemiology, 2015, Guidelines for Good Pharmacoepidemiology Practices (GPP), accessed June 2, 2025, https://www.pharmacoepi.org/resources/policies/guidances-08027/.

6. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance, 2023, ENCePP Guide on Methodological Standards in Pharmacoepidemiology, accessed June 2, 2025, https://encepp.europa.eu/encepp-toolkit/methodological-guide_en.

7. Guidance for industry *E8(R1) General Considerations for Clinical Studies* (April 2022), https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

8. Gatto NM, Campbell UB, Rubinstein E, Jaksa A, Mattox P, Mo J, and R Reynolds, 2022, The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework, Clin Pharmacol Ther, 111(1):122-134.

9. Brown JP, Hunnicutt JN, Ali MS, Bhaskaran K, Cole A, Langan SM, Nitsch D, Rentsch CT, Galwey NW, Wing K, and IJ Douglas, 2024, Quantifying Possible Bias in Clinical and Epidemiological Studies With Quantitative Bias Analysis: Common Approaches and Limitations, BMJ, 385:e076365.

10. Brown JP, Hunnicutt JN, Ali MS, Bhaskaran K, Cole A, Langan SM, Nitsch D, Rentsch CT, Galwey NW, Wing K, and IJ Douglas, 2024, Core Concepts in Pharmacoepidemiology: Quantitative Bias Analysis, Pharmacoepidemiol Drug Saf, 33(10):e70026.

11. Fox MP and TL Lash, 2020, Quantitative Bias Analysis for Study and Grant Planning, 2020, Ann Epidemiol, 43:32-36.

12. Fox MP, MacLehose RF, and TL Lash, 2021, Applying Quantitative Bias Analysis to Epidemiologic Data, Second Edition, Springer Cham.

13. Ding P and TJ VanderWeele, 2016, Sensitivity Analysis Without Assumptions, Epidemiology, 27(3):368-377.

14. Schneeweiss S, 2006, Sensitivity Analysis and External Adjustment for Unmeasured Confounders in Epidemiologic Database Studies of Therapeutics, Pharmacoepidemiol Drug Saf, 15(5):291-303.

15. Shi X, Liu Z, Zhang M, Hua W, Li J, Lee JY, Dharmarajan S, Nyhan K, Naimi A, Lash TL, Jeffery MM, Ross JS, Liew Z, and JD Wallach, 2024, Quantitative Bias Analysis Methods for Summary-Level Epidemiologic Data in the Peer-Reviewed Literature: A Systematic Review, J Clin Epidemiol, pub ahead of print Aug. 27, 2024, doi: 10.1016/j.jclinepi.2024.111507.

16. VanderWeele TJ and P Ding, 2017, Sensitivity Analysis in Observational Research: Introducing the E-Value, Ann Intern Med, 167(4):268-274.

17. Covvey JR, McClendon C, and MR Gionfriddo, 2024, Back to the Basics: Guidance for Formulating Good Research Questions, Res Social Adm Pharm, 20(1):66-69.

18. Wang SV, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin JA, Bartels DB, Kahler KH, Bessette LG, and S Schneeweiss, 2021, STaRT-RWE: Structured Template for Planning and Reporting on the Implementation of Real World Evidence Studies, BMJ, 372:m4856.

19. Gatto NM, Reynolds RF, and UB Campbell, 2019, A Structured Preapproval and Postapproval Comparative Study Design Framework to Generate Valid and Transparent Real-World Evidence for Regulatory Decisions, Clin Pharmacol Ther, 106(1):103-115.

20. Gatto NM, Vititoe SE, Rubinstein E, Reynolds RF, and UB Campbell, 2023, A Structured Process to Identify Fit-for-Purpose Study Design and Data to Generate Valid and Transparent Real-World Evidence for Regulatory Uses, Clin Pharmacol Ther, 113(6):1235-1239.

21. Strom BL, Kimmel SE, and S Hennessy, editors, 2019, Pharmacoepidemiology, Sixth Edition, Wiley-Blackwell.

22. Wang SV and S Schneeweiss, 2022, A Framework for Visualizing Study Designs and Data Observability in Electronic Health Record Data, Clin Epidemiol, 14:601-608.

23. Hallock H, Marshall SE, 't Hoen PAC, Nygard JF, Hoorne B, Fox C, and S Alagaratnam, 2021, Federated Networks for Distributed Analysis of Health Data, Front Public Health, 9:712569.

24. European Medicines Agency, 2021, Guideline on Registry-Based Studies - Scientific Guideline, accessed June 2, 2025, https://www.ema.europa.eu/en/guidance-registry-based-studies-scientific-guidance.

25. Guidance for industry *Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products* (December 2023) https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

26. Hernán MA and JM Robins, 2020, Causal Inference: What If, Boca Raton (FL): Chapman & Hall/CRC.

27. Weinstein EJ, Ritchey ME, and Lo Re V, 3rd, 2023, Core Concepts in Pharmacoepidemiology: Validation of Health Outcomes of Interest Within Real-World Healthcare Databases, Pharmacoepidemiol Drug Saf, 32(1):1-8.

28. European Medicines Agency, 2024, HMA-EMA Catalogues of Real-World Data Sources and Studies, accessed June 2 2025, https://catalogues.ema.europa.eu/.

29. International Committee of Medical Journal Editors (ICMJE), 2025, Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals, accessed Jun 2, 2025, https://www.icmje.org/recommendations/.

30. Guidance for industry *E6(R3) Good Clinical Practice* (September 2025), https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

31. Huybrechts KF, Bateman BT, and S Hernandez-Diaz, 2019, Use of Real-World Evidence From Healthcare Utilization Data to Evaluate Drug Safety During Pregnancy, Pharmacoepidemiol Drug Saf, 28(7):906-922.

32. Huybrechts KF, Bateman BT, and S Hernandez-Diaz, 2025, Modern Evidence Generation on Medication Effectiveness and Safety During Pregnancy: Study Design Considerations, Clin Pharmacol Ther, 117(4):895-909.

33. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance, 2023, ENCePP Guide on Methodological Standards in Pharmacoepidemiology, Annex 2: Guidance on Methods for the Evaluation of Medicines in Pregnancy and Breastfeeding, accessed June 2, 2025, https://encepp.europa.eu/encepp-toolkit/methodological-guide/annex-2-guidance-methods-evaluation-medicines-pregnancy-and-breastfeeding_en.

34. Council for International Organizations of Medical Sciences (CIOMS), 2021, Cumulative Pharmacovigilance Glossary: Version 1.1, accessed June 2, 2025, https://cioms.ch/cioms-cumulative-glossaries/.