Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document or the RealWorld Evidence Program, please email CDERMedicalPolicy-RealWorldEvidence@fda.hhs.gov

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)
Oncology Center of Excellence (OCE)

September 2021
Real World Data/Real World Evidence (RWD/RWE)

Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

Additional copies are available from:
Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353
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
10903 New Hampshire Ave., Bldg. 71, Room 3128
Silver Spring, MD 20993-0002
Phone: 800-835-4709 or 240-402-8010
Email: ocod@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)
Oncology Center of Excellence (OCE)

September 2021
Real World Data/Real World Evidence (RWD/RWE)

Contains Nonbinding Recommendations

Draft — Not for Implementation

TABLE OF CONTENTS

Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products
Guidance for Industry[1]

This draft guidance, when finalized, will represent 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 staff responsible for this guidance as listed on the title page.

I. INTRODUCTION AND SCOPE

The 21st Century Cures Act (Cures Act),[2] signed into law on December 13, 2016, is intended to accelerate medical product development and bring innovations faster and more efficiently to the patients who need them. Among other provisions, the Cures Act added section 505F to the Federal Food, Drug, and Cosmetic Act (FD&C Act) (21 U.S.C. 355g). Pursuant to this section, FDA created a framework for a program to evaluate the potential use of real-world evidence (RWE) to help support the approval of a new indication for a drug[3] already approved under section 505(c) of the FD&C Act or to help to support or satisfy postapproval study requirements (RWE Program).[4]

FDA is issuing this guidance as part of its RWE Program and to satisfy, in part, the mandate under section 505F of the FD&C Act to issue guidance about the use of RWE in regulatory decision-making.[5] The RWE Program will cover clinical studies that use real-world data (RWD) sources, such as information from routine clinical practice, to derive RWE.

[1] This guidance has been prepared by the Center for Drug Evaluation and Research (CDER) in cooperation with the Center for Biologics Evaluation and Research (CBER) and Oncology Center for Excellence (OCE) at the Food and Drug Administration.

[2] Public Law 114-255

[3] For the purposes of this guidance, all references to *drugs* include both human drugs and biological products. This guidance does not apply to medical devices. For information on medical devices, see guidance titled "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices" available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices.

[4] See *Framework for FDA's Real-World Evidence Program*, available at https://www.fda.gov/media/120060/download. The framework and RWE Program also cover biological products licensed under the Public Health Service Act.

[5] See section 505F(e) of the FD&C Act.

This guidance is intended to provide sponsors, researchers, and other interested stakeholders with considerations when proposing to use ***electronic health records***[6] (EHRs) or ***medical claims data*** in clinical studies[7] to support a regulatory decision on effectiveness or safety.

For the purposes of this guidance, FDA defines RWD and RWE as follows:[8]

- RWD are data relating to patient health status or the delivery of health care routinely collected from a variety of sources.

- RWE is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD.

Examples of RWD include data derived from EHRs, medical claims data, data from product and disease registries, patient-generated data including from in-home use, and data gathered from other sources that can inform on health status, such as digital health technologies. This guidance focuses on health-related data recorded by providers that can be extracted from two sources: EHRs and medical claims data. EHRs and medical claims data are types of ***electronic health care data*** that contain patient health information, and these data are widely used in safety studies and increasingly being proposed for use in effectiveness studies. EHR and medical claims data can be considered as data sources in various clinical study designs.

This guidance discusses the following topics related to the potential use of EHRs and medical claims in clinical studies to support regulatory decisions:

1. Selection of data sources that appropriately address the study question and sufficiently characterize study populations, exposure(s), outcome(s) of interest, and key ***covariates***

2. Development and ***validation*** of definitions for study design elements (e.g., exposure, outcomes, covariates)

3. Data ***provenance*** and quality during ***data accrual***, ***data curation***, and into the final study-specific dataset

This guidance does not provide recommendations on choice of study design or type of statistical analysis, and it does not endorse any type of data source or study methodology. For all study designs, it is important to ensure the reliability and relevance of the data used to help support a

[6] See the Glossary (section VII) for definitions of words and phrases that are in ***bold italics*** at first mention throughout this guidance.

[7] For the purposes of this guidance, the term *clinical studies* refers to all study designs, including, but not limited to, interventional studies where the treatment is assigned by a protocol (e.g., randomized or single-arm trials, including those that use RWD as an external control arm) and noninterventional studies where treatment is determined in the course of routine clinical care—i.e., observational studies (e.g., case-control or cohort studies). Throughout the guidance, FDA uses the terms *clinical studies*, *studies*, and *study* interchangeably.

[8] See *Framework for FDA's Real-World Evidence Program*, available at https://www.fda.gov/media/120060/download.

regulatory decision. For the purposes of this guidance, the term *reliability* includes data ***accuracy***, ***completeness***, provenance, and ***traceability***. The term *relevance* includes the availability of key ***data elements*** (exposure, outcomes, covariates) and sufficient numbers of representative patients for the study.

The contents of this document do not have the force and effect of law and are not meant to bind the public in any way, unless specifically incorporated into a contract. This document is intended only to provide clarity to the public regarding existing requirements under the law. FDA guidance documents, including this guidance, should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in FDA guidances means that something is suggested or recommended, but not required.

II. BACKGROUND

The FDA guidance for industry and FDA staff *Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data* (May 2013) focuses on the use of electronic health care data in pharmacoepidemiologic safety studies.[9] The 2013 guidance includes recommendations for documenting the design, analysis, and results of pharmacoepidemiologic safety studies to optimize FDA's review of protocols and study reports that are submitted to FDA.

This guidance complements the 2013 guidance by expanding on certain aspects of that guidance relating to the selection of data sources and also provides additional guidance for evaluating the relevance and reliability of both EHRs and medical claims data for use in a clinical study. This guidance also provides a broader overview of considerations relating to the use of EHR and medical claims data in clinical studies more generally, including studies intended to inform FDA's evaluation of product effectiveness.

III. GENERAL CONSIDERATIONS

For all studies using EHRs or medical claims data that will be submitted to FDA to support a regulatory decision, sponsors should submit protocols and statistical analysis plans before conducting the study. Sponsors seeking FDA input before conducting the study should request comments or a meeting to discuss the study with the relevant FDA review division. All essential elements of study design, analysis, conduct, and reporting should be predefined, and, for each study element, the protocol and final study report should describe how that element was ascertained from the selected RWD source, including applicable validation studies. More information about study elements is provided in Section V, Study Design Elements.

This guidance provides recommendations on selecting data sources to maximize the completeness and accuracy of the data derived from EHRs and medical claims for clinical studies. The use of certain study design features or specific analyses to address misclassified or missing information, as well as methods to achieve covariate balance, will be discussed in other

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

FDA RWE guidances focused on study design and analysis. This guidance addresses issues that are essential to determining the reliability and relevance of the data and that should be addressed in the protocol, including:

1. The appropriateness and potential limitations of the data source for the study question and to support key study elements.

2. Time periods for ascertainment of study design elements.

3. ***Conceptual definitions*** and ***operational definitions*** for study design elements (e.g., inclusion/exclusion criteria for study population, exposure, outcomes, covariates) and the results of validation studies. See Section V, Study Design Elements, for examples of conceptual and operational definitions for study design elements.

4. Quality assurance and quality control (QA/QC) procedures for data accrual, curation, and transformation into the final study-specific dataset.

IV. DATA SOURCES

Protocols submitted to FDA should identify all data sources proposed for the study, as well as other relevant descriptive information (discussed below). FDA does not endorse one data source over another or seek to limit the possible sources of data that may be relevant to answering study questions.

Each data source should be evaluated to determine whether the available information is appropriate for addressing a specific study hypothesis. Because existing electronic health care data were not developed to support regulatory submissions to FDA, it is important to understand their potential limitations when they are used for that purpose. Examples of potential limitations include:

1. The purpose of medical claims data is to support payment for care; claims may not accurately reflect a particular disease, or a patient may have a particular disease or condition that is not reflected in claims data.

2. EHR data are generated for use in clinical care and may also serve as a basis for billing and for auditing of practice quality measures. Data recorded in an EHR system depend on each health care system's practices for patient care and the clinical practices of its providers. In addition, data collection is limited to the data captured within an EHR system or network, and may not represent comprehensive care (e.g., care obtained outside of the health care system).

3. For prospective clinical studies proposing to use EHRs, it may be possible to modify the EHR system for the purpose of collecting additional patient data during routine care through an add-on module to the EHR system. However, given the limited ability to add

modules to collect extensive additional information, EHR-based data collection may still not be comprehensive.

The historical experience with and use of the selected data source for research purposes should be described in the protocol. This description should include how well the selected data source has been shown to capture study elements (e.g., inclusion and exclusion criteria, exposures, outcomes, key covariates) and how the data can be validated for a particular research activity.

A. Relevance of the Data Source

There are differences in the practice of medicine around the world and between health care systems that may affect the relevance of the data source to the study question. Patients in different types of commercial or government health care payment programs can differ in a range of characteristics, such as age, socioeconomic status, health conditions, risk factors, and other potential ***confounders***. Various factors in health care systems and insurance programs, 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 health care system might differ in disease severity, or other disease characteristics, from patients on the same therapy in another health care system. It is also important to identify whether the data sources cover all populations relevant to the study if those sources are to be used to examine the study hypothesis.

FDA recommends providing:

1. The reason for selecting the particular data sources to address the specific hypotheses.

2. Background information about the health care system, including (if available) any specified method of diagnosis and preferred treatments for the disease of interest, and the degree to which such information is collected and validated in the proposed data sources.

3. A description of prescribing and use practices in the health care system (if available), including for approved indications, formulations, and doses.

For non-U.S. data sources, FDA recommends providing an explanation of how all of these factors might affect the generalizability of the study results to the U.S. population.

B. Data Capture: General Discussion

A record in EHR systems or medical claims databases is generated only if there is an interaction of the patient with the health care system. Because EHR and medical claims data are collected during routine care and not according to a prespecified research protocol, information needed to address certain questions in a proposed study may not be present in EHR and medical claims data sources. Sponsors should demonstrate that each data source contains the detail and completeness needed to capture the study populations, exposures, key covariates, outcomes of interest, and other important parameters (e.g., timing of exposure, timing of outcome) that are relevant to the study question and design.

1. Enrollment and Comprehensive Capture of Care

Continuity of coverage (enrollment and disenrollment) should be addressed when using EHR and medical claims data sources, given that patients often enroll and disenroll in different health plans when they experience changes in employment or other life circumstances. The validity of findings from a study using these data depends in part on the documentation of the migration of patients into and out of health plans and health care delivery organizations. Such documentation allows the definition of enrollment periods (during which data are available on the patients of interest) and disenrollment periods (when data are not available on patients). Definitions of *enrollment* or *continuous coverage* should be developed and documented in the protocol.

In addition, FDA recommends addressing the comprehensiveness of the data sources in capturing aspects of care and outcomes that are relevant to the study question. This information will help evaluate the likelihood that all exposures and outcomes of interest will be captured for regulatory decision-making. For example, outpatient data sources that do not include hospitalization data would generally not be appropriate for studying outcomes likely to result in hospitalization. A second example is a study where an outcome is dependent on a specific frequency of laboratory tests, and clinicians do not typically order those tests at such a frequency.

FDA recommends specifying how all relevant populations, exposures, outcomes, and covariates will be captured during the ***study period***, particularly in situations where data availability varies greatly over time. The data sources should contain adequate numbers of patients with adequate length of follow-up to ascertain outcomes of interest based on the biologically plausible time frame when the outcome, if associated with the exposure, might be expected to occur. Information should be provided about the distribution of length of follow-up for patients in the data sources because the length of follow-up may inform whether the selected data sources are adequate or whether additional supportive data are needed to ascertain long-latency outcomes.

In general, EHR and medical claims data do not systematically capture the use of nonprescription drugs or drugs that are not reimbursed under health plans, or immunizations offered in the workplace. If these exposures are particularly relevant to the study question, the data source may not be suitable, or the protocol should describe how this information gap will be addressed.

Obtaining comprehensive drug coverage and medical care data on patients with certain types of privacy concerns (e.g., sexually transmitted infection, substance abuse, mental health conditions) can be challenging and failure to do so can result in incomplete or erroneous information. Patients with these conditions may receive treatment in federally qualified health centers, or in private clinics where an insurance claim may not be generated if self-payment is used. In addition, certain populations more often enroll in experimental clinical trials—e.g., patients with certain cancers or patients who receive their medications under pharmaceutical company assistance programs. In such cases, patients' health data may not be fully captured in most electronic health care data sources. If these issues are relevant to the study question of interest, the protocol should describe how the issues will be addressed.

2. Data Linkage and Synthesis

Data linkages can be used to increase the breadth and depth of data on individual patients over time and provide additional data for validation purposes. If the study involves establishing new data linkages between internal data sources (e.g., mother-infant linkages) or external data sources (e.g., vital records, disease and product registries, biobank data), the protocol should describe each data source, the information that will be obtained, linkage methods, and the accuracy and completeness of data linkages over time. If the study involves generating additional data (e.g., interviews, mail surveys, computerized or mobile-application questionnaires, measurements through digital health technologies), the protocol should describe the methods of data collection and the methods of integrating the collected data with the electronic health care data. Probabilistic and deterministic approaches to data linkage may result in different linkage quality, albeit both approaches can have value depending on the scenario. The deterministic approach for data linkage uses records that have an exact match to a unique or set of common identifiers, and the match status can be determined using a single or multiple step process. The probabilistic approach for data linkage uses less restrictive steps in which the identifiers compared consist of fewer variables or part of them (Carreras et al., 2018). When a probabilistic approach is used, the analysis plan should include testing the impact of the degree of match and robustness of findings. See Section VI, Data Quality During Data Accrual, Curation, and Transformation into the Final Study-Specific Dataset.

For studies that require combining data from multiple data sources or study sites, FDA recommends demonstrating whether and 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.

Because patients typically visit multiple health care sites, especially in geographically contiguous areas, the inclusion of de-identified data from many sites creates the possibility that there will be multiple records from different health care sites for a single individual. The existence of multiple records of the same person in different sites can result in overcounts of a particular data measure or, alternatively, if some site records are not available, can result in a collection of patient histories that reflect only a fraction of the patient's total health care history. Specific attention to data curation including individual level and population level linkages and understanding of many-to-one and 1:1 linkage is fundamental to assessing the appropriateness of a new data linkage. This scenario is not an issue with data sources that share a unique patient identifier across all sites (e.g., a multi-site hospital network) and only occurs if the patient seeks care outside the network. FDA recommends considering and documenting the type of curation performed to address duplication or fragmentation issues and documenting approaches taken to address issues that cannot be fully rectified by curation. See Section VI, Data Quality During Data Accrual, Curation, and Transformation into the Final Study-Specific Dataset.

3. Distributed Data Networks

Distributed data networks (or systems) of EHRs and medical claims data systems, often combined with the use ***Common Data Models*** (CDMs), have been increasingly used for medical product safety surveillance and research purposes. The primary benefit of using a distributed network in which data from multiple sites are transformed into a single CDM, is the ability to execute an identical query (without any or substantial modifications) on multiple datasets. In some distributed data networks, queries can be run simultaneously at all network sites or at each site asynchronously, with results combined at a coordinating center for return to the end user. There are a number of the commonly used operational models employed by distributed data networks. Some networks are managed by a single business entity using a consistent EHR system or medical claims database structure and while data are maintained at many locations, they are structured and managed in a consistent manner (e.g., the U.S. Department of Veterans Affairs Veterans Health Administration). Another approach is a hybrid distributed model in which a subset of data from many remote sites is sent to a centralized repository to allow direct research on a combined data set (e.g., U.S. Centers for Disease Control and Prevention's National Syndromic Surveillance System, previously known as BioSense 2.0). A third commonly used approach is seen in networks of data systems with multiple owners and database structures, with data structured and managed differently from location to location (e.g., the member sites of FDA's Sentinel system). In this model, research questions are sent to the various network member sites and answers returned to a central location for collation and reporting.

The latter type of networks, comprised of disparate data systems such as the Sentinel system, are facilitated by the use of CDMs. Networks using CDMs also typically provide tools and methodologies for analysis, a consistent level of data curation, and periodic revision of the data model to incorporate new data concepts as needed by researchers. Additionally, methodologies have been developed that allow the ability to translate data from one CDM to another, however these involve additional data transformations, which present added quality considerations. Data curation and transformation into a CDM, as well as general QA/QC processes, are discussed in Section VI, Data Quality During Data Accrual, Curation, and Transformation into the Final Study-Specific Dataset.

Distributed data networks are typically comprised of EHR, medical claim, or registry data. However, combining many data sources, especially with the addition of data transformation into a CDM, adds a layer of complexity that should be considered. Because there are many different configurations of distributed health data networks, the configurations discussed in this guidance should not be considered comprehensive.

Transforming disparate database structures into a common health network with a CDM allows research across health care sites that would otherwise be more complex and costly. However, CDMs can introduce additional challenges for researchers to consider. Many CDMs, including those developed for FDA's Sentinel system, Biologics Effectiveness and Safety Initiative, and the National Patient-Centered Clinical Research Network, were created to satisfy a specific set of research purposes; the choice of data captured in a CDM is optimized for the types of data measures and detail needed for the intended use (e.g., Sentinel system for postmarket safety surveillance to inform regulatory decision-making, the National Patient-Centered Clinical Research Network for patient-centered outcomes research). Therefore, data in CDM-driven

networks rarely contain all of the source information present at the individual health care sites, and the data elements chosen for a given CDM network may not be sufficient for all research purposes or questions. Furthermore, CDMs typically often have many data elements within the model that are optional—that is, although the model has such data elements available to be filled with data, the individual sites can choose whether to put their original data into the optional fields.

Before using a CDM-driven network, data elements collected by the CDM should be considered—including whether needed data elements exist in the model and, if so, whether they are required or optional elements—to determine suitability for the study and whether identified deficiencies can be addressed by supplementing with customized study-specific data elements, collecting additional data, or using other data elements present in the dataset that are reasonable proxies for the missing information. It should be noted, such workarounds would involve additional considerations by the sponsor such as the work involved with validating proxy endpoints or any human subject research considerations that involve additional data. Suitability may also be improved with more flexible CDMs that are frequently expanded for new uses. For information on proxy variables, see Section IV.C, Missing Data: General Considerations.

4. Computable Phenotypes

Standardized ***computable phenotypes*** can facilitate identification of similar patient populations and enable efficient selection of populations for large-scale clinical studies across multiple health care systems. A computable phenotype definition should include metadata and supporting information about the definition, its intended use, the clinical rationale or research justification for the definition, and data assessing validation in various health care settings (Richesson et al. 2016). The computable phenotype definition, composed of data elements and phenotype algorithm, should be described in the protocol and study report and should also be available in a computer-processable format. Clinical validation of the computable phenotype definition should be described in the protocol and study report. For additional information on validation, see Section IV.D, Validation: General Considerations.

5. Unstructured Data

Large amounts of key clinical data are unstructured data within EHRs, either as free text data fields (such as physician notes) or as other non-standardized information in computer documents (such as PDF-based radiology reports). To enhance the efficiency of data abstraction, a range of approaches, including both existing and emerging technologies, are increasingly being used to convert unstructured data into a computable format. More recent innovations include technology-enabled abstraction whereby software provides a mechanism for human data abstractors (e.g., tumor registrars) to do their work in a consistent and scalable fashion.

Technological advances in the field of ***artificial intelligence*** (AI) may permit more rapid processing of unstructured electronic health care data. Advances include natural language processing, machine learning, and particularly deep learning to: (1) extract data elements from unstructured text in addition to structured fields in EHRs; (2) develop computer algorithms that

identify outcomes; or (3) evaluate images or laboratory results. FDA does not endorse any specific AI technology.

All of these methods are computer-assisted to various levels but currently require a significant amount of human-aided curation and decision-making, injecting an additional level of data variability and quality considerations into the final study-specific dataset. If the protocol proposes to use AI or other derivation methods, the protocol 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. Relevant impacts on data quality should be documented in the protocol and analysis plan.

C. Missing Data: General Considerations

There are two broad cases in which information may be absent from the data sources. The first case is when the information was intended to be collected (e.g., structured field present in the EHR), but is absent from the data sources. This is an example of traditional ***missing data***. The second case is when the information was not intended to be collected in the EHR and medical claims data and is therefore absent. It is important to distinguish between these two cases and understand the reasons why information is present or absent in EHRs and medical claims. For example, lack of information about the result of a laboratory test could be caused by different circumstances: (1) the test might not have been ordered by the health care provider; (2) the test might have been ordered but not conducted; (3) the test might have been performed, but the result was not stored or captured in the data source; or (4) the test might have been performed 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. Because providers might order a laboratory test based on a patient's characteristics, the decision not to order the test or a patient's decision to forgo the test may have implications on the data's fitness for use in a proposed study.

As discussed above, data linkage is one way to address missing data. It may also be possible to identify a variable that is a proxy for the missing data. An example of a potential proxy variable includes low-income subsidy under the Medicare Part D prescription drug program as a proxy for a patient's socioeconomic status.

The protocol and the statistical analysis plan should be developed and based on an understanding of reasons for the presence and absence of information. Descriptive analyses should be included to characterize the missing data. Assumptions regarding the missing data (e.g., missing at random, missing not at random) underlying the statistical analysis for study end points and important covariates should be supported and the implications of missing data considered.

D. Validation: General Considerations

Studies using EHR and medical claims data sources should include conceptual definitions for important study variables, including study population inclusion and exclusion criteria, exposure, outcome, and covariates. A conceptual definition should reflect current medical and scientific

thinking regarding the variable of interest, such as: (1) clinical criteria to define a condition for population selection or as an outcome of interest or a covariate; or (2) measurement of drug intake to define an exposure of interest.

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 EHR or medical 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.

Because operational definitions are usually imperfect and cannot accurately classify the variable of interest for every subject, a resulting ***misclassification*** can lead to false positives and false negatives (Table 1) and may bias the association between exposure and outcome in a certain direction and degree. Although complete verification[10] of a variable of interest minimizes misclassification and maximizes study internal validity, understanding the implications of potential misclassification for study internal validity and study inference is the key step in determining what variables of interest might require validation and to what extent. For example, in a study to quantify a drug effect, internal validity should be optimized, and misclassification of key variables should be minimized to accurately measure the association. Some misclassification might be tolerable in some studies when the presence of misclassification is not expected to change the interpretation of results (e.g., for signal detection, or when the hypothesized effect size is large and the impact of misclassification on the measure of association is deemed minimal).

To understand how potential misclassification of a variable of interest (e.g., exposure, outcome, covariate) might impact the measure of association and the interpretation of results, sponsors should consider: (1) the degree of misclassification; (2) differential versus non-differential misclassification (e.g., differential misclassification of outcome by exposure); (3) dependent versus independent misclassification (e.g., correlated misclassifications of exposure and outcome when both are self-reported in the same survey); and (4) the direction toward which the association between exposure and outcome might be biased.

Although complete verification of a study variable is considered the most rigorous approach, there are scenarios where verifying a variable for every subject might not be feasible (e.g., a very large study population, lack of reference standard[11] data for all study subjects) and assessing the performance of the variable's operational definition might suffice. Based on the performance measures described in Table 1, sponsors should consider whether validating the variable to a

[10]For the purposes of this guidance, complete verification involves assigning an accurate value to the variable of interest for each study subject based on a reference standard of choice. For example, medical record review can be used in conjunction with a conceptual definition to determine whether a subject meets a critical inclusion criterion or has experienced the outcome event. (To a variable extent, adjudication may be involved in this process.)

[11] For purposes of this guidance, reference standard is the best available benchmark, also referred to as "gold standard."

greater extent (e.g., all positives classified by the operational definition) is necessary and discuss with the relevant review division.

Because the performance of an operational definition is dependent on various factors, such as data source, study population, study time frame, and choice of reference standard, FDA recommends assessing the performance of operational definitions in an adequately large sample of the study population as part of the proposed study, using justified sampling methods (e.g., random sampling, stratified sampling). If sponsors propose to use an operational definition that has been assessed in a prior study, ideally those operational definitions assessed in the same data source and in a similar study population should be selected. In addition, secular trends in disease, diagnosis, and coding may necessitate assessment of the operational definition using more recent data. The quality of prior studies used to establish ***sensitivity***, ***specificity***, and predictive values should always be evaluated.

The protocol should include a detailed description of the planned validation, including justification for the choice of a reference standard, validation approach, methods, processes, and sampling strategy (if applicable). If a previously assessed operational definition is proposed, additional information should be provided, including in what data source and study population and during what time frame the assessment was conducted, the value of the assessed performance measures, and a discussion of whether the performance measures are applicable to the proposed study. FDA also recommends including in the protocol prespecified sensitivity analyses to demonstrate whether and how bias, if present, might impact study findings based on the validation data.

For further discussion about the validation of study design elements, see Section V.C.5, Validation of Exposure; Section V.D.3, Validation of Outcomes; and Section V.E.3, Validation of Confounders and Effect Modifiers.

Table 1: Schematic Representation of the Calculation of Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) for a Binary Variable

Condition based on proposed operational definition	Condition based on reference standard: Yes	Condition based on reference standard: No	Total	
Yes	a (true positive)	b (false positive)	a+b	PPV = a/(a+b)
No	c (false negative)	d (true negative)	c+d	NPV = d/(c+d)
Total	a+c	b+d	N	
	Sensitivity = a/(a+c)	Specificity = d/(b+d)		

V. STUDY DESIGN ELEMENTS

The ascertainment and validation of key study design elements are discussed in detail below. The study questions of interest should be established first, and then the data source and study design most appropriate for addressing these questions should be determined. The study should not be designed to fit a specific data source, because the limitations of a specific data source may restrict the options for study design and limit the inferences that can be drawn. Considerations regarding study design and analysis when using RWD sources will be discussed in other RWE guidance documents.

A. Definition of Time Periods

FDA recommends clearly defining the various time periods pertinent to the study design in the protocol (e.g., time periods for identifying study population, defining inclusion and exclusion criteria, assessing exposure, assessing outcome, assessing covariates, following up with patients). The focus of the time scale (e.g., calendar time, age, time since exposure) should be explicitly described with adequate detail on data availability of the time unit (e.g., year, month, day, hour, minute) required to answer the study question.

The protocol should justify proposed time periods and the potential impact on study validity. For example, justification should be provided regarding whether the time period before exposure is appropriate for identifying the study population and the important baseline covariates, whether the follow-up time is sufficient for observing the occurrence of study outcomes, and whether the time period for updating information on time-dependent covariates is suitable to capture the changes of those variables. In addition, when considering outcome definitions, disease onset (e.g., early symptoms) may need to be distinguished from a confirmed diagnosis, as appropriate to the study question. When defining the beginning and the end of the follow-up time for outcome assessment, consider the biologically plausible time frame when the outcome, if associated with the exposure, might be expected to occur.

The protocol should also address potential temporal changes in the standard of care, the availability of other treatments, diagnosis criteria, and any other relevant factors that are pertinent to the study question and design. Other relevant factors may include insurance formulary changes (if known), step therapy, and laboratory assay changes. Before developing the study approach, sponsors should discuss with the relevant FDA review division the capability of data to capture such potential temporal changes and the impact of the potential temporal changes on internal validity.

B. Selection of Study Population

The protocol should include a detailed description of methods for determining how inclusion and exclusion criteria (e.g., demographic factors, medical condition, disease status, severity, biomarkers) will be implemented to identify appropriate patients meeting these criteria from the data source. The protocol should address the completeness and accuracy of the information collected in the proposed data source to fulfill the inclusion and exclusion criteria.

Key variables used to select the study population should be validated. For example, to assess the drug effect in patients with immune thrombocytopenic purpura, the disorder ascertained by operational definition International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis code 287.31 should be validated based on the conceptual definition of the disorder, which includes signs and symptoms, levels of platelets, and exclusion of other possible causes of thrombocytopenia.

In certain circumstances, key variables (e.g., gestational age for pregnancy studies) required to fulfill the inclusion and exclusion criteria may be generated by the health care provider using information available at the point of care. For example, health care providers may enter the calculated gestational age in an EHR based on patient self-reported last menstrual period, ultrasound dating, and other relevant information. If such data are used, the protocol should describe the source of information and the methods health care providers use to generate the data (if known).

C. Exposure Ascertainment and Validation

Considerations discussed in this section regarding exposure ascertainment in medical claims data or EHRs primarily apply to noninterventional studies, given that the assignment of exposure is documented in interventional studies.

1. Definition of Exposure

For the purposes of this guidance, the term *exposure* applies to the medical product or regimen of interest being evaluated in the proposed study. The product of interest is referred to as *the treatment*, and may be compared to no treatment, a placebo, standard of care, another treatment, or a combination of the above. Other variables that could affect the study outcome are considered covariates and are discussed in Section V.E, Covariate Ascertainment and Validation. The exposure definition should include information about the drug dose, formulation, strength, route, timing, frequency, and duration the product studied (if relevant). It may also be necessary to describe the specific manufacturer of a product (e.g., when a proper name for a vaccine is used by different manufacturers).

The description of exposure should include the intended or prescribed use of the product (e.g., the number, frequency, and specific doses), the period between initiation of exposure and the earliest time one might reasonably expect to see an effect, and the expected duration of effect. This will usually require an understanding of the pharmacological properties of the drug—for example, that a one-time infusion to prevent osteoporosis may have an effect for several months. See Section V.C.3, Ascertainment of Exposure: Duration, and Section V.C.4, Ascertainment of Exposure: Dose.

2. Ascertainment of Exposure: Data Source

Sponsors should be able to demonstrate an ability to identify the specific products of interest in the proposed data source, demonstrating that the data source contains data fields and codes that allow identification of the specific products of interest (e.g., through specific coding). For

example, it is not always possible to infer a specific vaccine formulation from the billing or diagnostic code alone, such as in systems where a single billing code is used for multiple vaccines. The protocol should describe the coding system used, the level of granularity represented (e.g., using RxNorm mapping to the National Drug Code [NDC] identifiers), and the specificity attained by the coding system.

When relying on coded data, the operational exposure definitions should be based on the coding system of the selected data source and reflect an understanding of the prescription, delivery, and reimbursement characteristics of the drug (if applicable) in that data source. For example, in the United States, the operational definition should include the appropriate pharmacy codes (NDC or Healthcare Common Procedure Coding System) to capture the use of the drug in various settings. This approach is particularly important in the case of non-oral drugs that may be assigned different codes depending on how they are obtained. For example, patients using an injectable drug can purchase it from the pharmacy, in which case the NDC code would be recorded, or it can be administered by the provider for the patient and the drug and its administration would be recorded using the HCPCS J code.[12]

It is also essential to report operational definitions and methods when combining information from unstructured and structured data. Emerging methods may involve review of unstructured information in medical records combined with pharmacy dispensing and physician prescribing data and notes to provide an assessment of whether a person was prescribed and received the medication of interest, as well as whether there are problems with the patient continuing the medication. An example of such methods is found in ascertainment of aspirin exposure in a retrospective cohort study of veterans undergoing usual care colonoscopy (Bustamente et al. 2019).

When using a medical claims data source, it is important to consider that there could be dispensed prescriptions that were not associated with insurance claims if these uncaptured prescriptions are relevant exposures for the study. Uncaptured prescriptions might include low-cost generic drugs, drugs obtained through discount programs, samples provided by pharmaceutical companies and dispensed by health care providers, and drugs sold via the internet or patient out-of-pocket purchases. In addition, nonprescription drugs and dietary supplements are not generally captured in electronic health care databases. It is important to address the likelihood of incomplete exposure ascertainment and its effect on study validity, see Section V.C.5, Validation of Exposure.

3. Ascertainment of Exposure: Duration

The data source should capture the relevant exposure duration (anticipated use of a product over time). Given that some medical products are designed as one-time exposures (e.g., vaccines), and other products may be intended for use over extended periods of time, the suitability of a data source will vary with the specific medical product under investigation. FDA recommends describing the duration of exposure as well as the period during which the exposure is having its

[12] A drug's J code is a Healthcare Common Procedure Coding System Level II code used in medical claims to report injectable drugs that ordinarily cannot be self-administered; chemotherapy, immunosuppressive drugs, and inhalation solutions; and some orally administered drugs.

effect relative to the outcome of interest. Duration may refer to continuous exposure or cumulative exposure, depending on the study question. For some products, an immediate or near-immediate effect is expected; for other products, an effect is expected after a time interval (e.g., drugs that promote bone strength). FDA recommends considering the duration of continued drug effect after treatment discontinuation to include the entire period in which the drug effect may occur. For example, a vaccine effect may persist for years after vaccination, and persons might be considered exposed during that period. On the other hand, an anticoagulant's effects would not extend beyond several hours or days. FDA also recommends justifying the units (e.g., hours, days) selected for estimating the duration of exposure and ensuring the data are available in those units.

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.[13] FDA recommends describing and justifying in the protocol how researchers will measure use, address potential gaps in therapy in the data source, and handle refill stockpiling if there are early refills. Intermittent therapies (e.g., drugs used to treat pain on an as-needed basis) and therapies for which samples are often provided to patients (e.g., expensive drugs, drugs that are new to the market) present challenges in accurately assessing the actual exposure and duration of exposure, see Section V.C.5, Validation of Exposure.

4. Ascertainment of Exposure: Dose

Data about exposure should include information about dose. Depending on the exposure and the question of interest in the study, it may be useful to describe the dose of each administration or a daily dose, as well as an estimated ***cumulative dose***.

It is reasonable to begin with the dose information provided in the data source, and then discuss in the protocol or study report the specific assumptions made when estimating the dose of the exposures of interest, especially for pediatric patients. See Section V.C.6, Dosing in Special Populations. It is also important to report how different dosage forms (e.g., parenteral versus oral) will figure into the dose calculation if multiple forms are available.

5. Validation of Exposure

Other than for medications administered in hospital settings or infusion settings, electronic health care data capture prescriptions of drugs and the dispensing of drugs to patients, but generally do not capture actual patient drug exposure because this depends on patients obtaining and using the prescribed therapy.

Validation ideally involves a comparison of the exposure classification in the proposed data source with a reference data source,[14] and produces estimates of misclassification that can be used in sensitivity analyses. Validation might begin with defining the conceptual and operational

[13] This guidance does not address issues related to medication adherence.

[14] In certain cases, the RWD source may be the only reference. For example, if exposure is defined by whether the patient paid for the prescription, medical claims data may be used, and this information will be the reference source.

definitions. For example, to define new use of drug X in a particular study, the conceptual definition may be "initiation of drug X and no exposure to drug X in the past 365 days," and the operational definition would be "at least one outpatient prescription claim for drug X (identified by NDC code xxx), and no claims for drug X in 365 days before the dispensing date of the prescription." For prescribed medications used in outpatient settings, dispensing or billing data would tend to be more accurate than most EHRs in reflecting exposure to a drug by documenting that the prescriptions were filled. In such cases, validation of EHR prescribing data by examining medical claims data may be warranted. For drugs administered in the health care setting (e.g., vaccines, injectables, blood products), administration recorded in the EHR may provide more complete information than is available in medical claims records. In these cases, it may be useful to validate medical claims data by examining the EHR. In certain situations, when reference data sources are not available, additional studies conducted in the same population or published in the literature can provide estimates of potential misclassification of exposure status (e.g., survey of study participants to assess intake of drug, published reports of numbers of people obtaining vaccinations through pharmacies/workplaces/schools).

FDA recommends documenting the methods used to calculate and validate duration, dose, switching, and other characteristics of exposure. Validation and misclassification issues should be addressed in appropriate study documents.

6. Dosing in Special Populations

In addition to reporting validated information about the dose prescribed, dispensed, or administered, additional information may be necessary to permit an assessment of whether dosing was appropriate for special populations (e.g., if there was significant underdosing or overdosing). For example, in assessing dosing in patients taking drugs with substantial renal clearance, it may be necessary to have access to measurements of serum creatinine, creatinine clearance, or estimated glomerular filtration rate to assess appropriateness of dosing. Another example is when estimating exposure in pediatric populations where it may be necessary to obtain the patient's weight and describe the dose within weight categories. The need for additional data to permit appropriate assessment of dosing may occur more frequently with claims data, but can also occur when using EHRs if necessary data are absent.

7. Other Considerations

Selecting an appropriate comparator is an essential part of a clinical study. The patients providing comparator data should be defined clearly and with adequate detail in the protocol. The protocol should discuss the reasoning for selecting the: (1) source of comparator data; and (2) the time period (if the comparator group is not concurrent with the treatment group). Because a comparator agent may differ from the product of interest in specific indication within a disease, contraindication, safety profile, or user's disease severity or comorbidity, as well as other patient characteristics, it is important to ensure adequate data are available for FDA to assess the comparability of the exposed and comparator populations.

Relevant ***concomitant medication*** use should be described and ascertained from the data source. A study's definition of concomitant medication use should be described in detail. Definitions of

concomitant medication use might include instances when drugs are dispensed on the same day, when drugs have overlapping days' supply, or when patients have filled prescriptions for two or more drugs during the study period. Limitations to ascertainment of concomitant drugs (e.g., nonprescription drugs) should also be described.

D. Outcome Ascertainment and Validation

A crucial step in selecting a data source is determining whether it captures the clinical outcome of interest. Because electronic health care data typically capture outcomes that are brought to the attention of a health care professional and documented in the medical record, outcomes representing mild symptoms or events occurring outside of medical care (e.g., out-of-hospital death) will not generally be well-captured. Conversely, discrete outcomes or acute events (e.g., stroke, myocardial infarction, new infection) are more likely to be captured than worsening of existing problems (e.g., depression, psoriasis, arthritis) that do not lead to discernible events. Unlike traditional clinical trials, studies exclusively using electronic health care data to ascertain outcomes likely do not have protocol-defined follow-up visits and may not have monitoring of events at intervals necessary for outcome ascertainment. In addition, the assessment of the outcome of interest is likely more standardized and comprehensive in traditional clinical trials. Therefore, the availability, accuracy, and completeness of data on the outcome of interest as well as the need for external data linkage should be carefully considered. Whether and to what degree a data source captures the outcome of interest should be assessed before study initiation and be independent of the exposure of interest.

1. Definition of Outcomes of Interest

Many outcomes involve diagnoses recorded by physicians as part of routine care. To minimize the effect of variability in practice by different physicians and over time (e.g., using different diagnosis and classification criteria, coding the same event in different ways), FDA recommends defining an outcome of interest based on the clinical, biological, psychological, and functional concepts of the condition, as appropriate. The conceptual definition for the outcome of interest (also referred to as the *case definition*) should reflect the medical and scientific understanding of the condition and might vary by study. For example, for anaphylaxis, the conceptual definition (or case definition) may include the following clinical criteria: sudden onset, rapid progression of signs and symptoms, ≥1 major dermatological criterion, and ≥ 1 major cardiovascular or respiratory criterion. The protocol should include a detailed description of the conceptual definition, including the signs, symptoms, and laboratory and radiology results that would confirm the outcome.

Conceptual definitions should be able to be operationalized in RWD sources. For example, randomized controlled trials in oncology typically use tumor-based outcomes of interest in the setting of specific timing and frequency of follow-up assessment and often include molecular or biomarker testing that may not be standard-of-care in the clinical practice settings. Since achievement of an objective response (tumor shrinkage), or the date of tumor progression based on standardized clinical trial criteria (e.g., RECIST 1.1) is not typically captured in RWD sources, proxy measures or multi component definitions may need to be explored and their use justified. In general, it may be easier to capture outcomes that have well-defined diagnostic

criteria that are likely to be consistently captured in RWD, such as stroke, myocardial infarction or pulmonary embolism, compared to outcomes that are more subjective or scaled in nature, such as worsening of joint pain in rheumatoid arthritis or worsening of depression symptoms. Sponsors should discuss proposed outcomes definitions with the FDA review division.

2. Ascertainment of Outcomes

To help identify potential cases in the selected data source and study population, operational definitions using diagnosis and procedure codes (e.g., ICD-9-CM, ICD-10), laboratory tests (e.g., LOINC) and values, or unstructured data (e.g., physician's encounter notes, radiology and pathology reports) should be developed based on the conceptual definition of the outcome of interest. If the operational definition includes information abstracted from unstructured data in the EHR or another data source (e.g., mention of spina bifida in birth certificate records for the identification of neural tube defects in infants), the protocol should provide a detailed description and rationale for the methods and tools used to process the unstructured data and the validation of those methods. See Section IV.B.5, Unstructured Data, for additional information on unstructured data. When patient- or physician-generated data (e.g., data required for subjective end points) are proposed to assess the outcome of interest or to complement operational definitions, the protocol should specify how the outcome measure (e.g., sign score, severity index) will be defined and constructed and validated, if applicable, and how the data will be collected.

The sensitivity and specificity of an operational definition are imperfect when there is outcome misclassification. Given that it is usually not possible for sensitivity and specificity to be perfect (i.e., 100%), outcome misclassification might result in both false positives and false negatives. FDA recommends considering the potential impact of outcome misclassification on study validity when developing or selecting an operational definition for the proposed study. For example, when studying infrequently occurring outcomes in a cohort study, given the low prevalence of the outcome event, it is important to achieve high specificity to minimize false-positive cases and high sensitivity so that more true cases can be captured.

Operational definitions developed for one data source or study population might not perform as well in other sources or populations, due to database-specific sensitivity and specificity as well as variable disease prevalence. ***Positive predictive value*** (PPV) and ***negative predictive value*** (NPV) are related to sensitivity and specificity and are a direct function of prevalence of the outcome in the population in which the predictive values are measured. Therefore, PPV and NPV are variable by data source and study population characteristics (e.g., demographic factors, underlying diseases, comorbidities, clinical settings).

The protocol should include a detailed description of the operational definition, the coding system, the rationale and associated limitations of information selected to construct the operational definition (e.g., selection of primary or secondary diagnosis codes for which the order may not correspond to their medical importance), and the potential impact on outcome misclassification. If the performance of the operational definition has been assessed in prior studies, the applicability to the proposed study should be discussed. Further, because the case definition used in prior studies to establish sensitivity, specificity, and predictive values might

include different diagnostic criteria from the conceptual definition developed for the proposed study, proper use of the performance measures assessed in prior studies should be carefully considered.

3. Validation of Outcomes

FDA expects validation of the outcome variable to minimize outcome misclassification. Although complete verification of the outcome variable is considered the most rigorous approach, there are scenarios where verifying outcome for every subject might not be feasible and assessing the performance of the operational definition of the outcome might suffice. Outcome validation involves using a clinically appropriate conceptual outcome definition to determine whether a patient's status, classified by an operational definition, truly represents the outcome of interest, typically by reviewing clinical details recorded in the patient's medical records in either electronic or paper format.

FDA recommends using standardized medical record review processes, including the use of standardized tools, documentation of process, and training of personnel. A standard and reproducible process is critical for minimizing intra- and inter-rater variability, especially for multi-site studies in which medical records usually cannot be shared across systems and a centralized medical record review is not possible. Even with a centralized medical record review, a standardized process helps to ensure that the same criteria are applied by different adjudicators or a single adjudicator over time. Reporting of comparison metrics (e.g., kappa statistic) is useful to ensure replicability. An estimated medical record retrieval rate should be justified in the protocol, and the implications for internal and external validity should be discussed. In addition, because knowledge of a patient's exposure status may influence the observer and result in differential misclassification, blinding of the abstractor and adjudicator to exposure status should be considered by masking the study question or redacting the exposure information, especially when the abstractor or adjudicator may associate the exposure with the outcome of interest. The protocol should provide a description of how observer bias will be handled.

Ideally, through complete verification of the outcome variable, each subject is assigned an accurate value of the outcome variable to minimize outcome misclassification and improve study internal validity. In practice, a more commonly used approach is to assess the performance of an operational definition in validation studies. Performance measures, such as sensitivity, specificity, and predictive values, do not accurately classify cases and non-cases; rather, they inform the degree of outcome misclassification and facilitate the interpretation of results in the presence of misclassification.

PPV is often assessed in validation studies. PPV is the proportion of potential cases identified by an operational definition that are true-positive cases. Therefore, PPV informs the degree to which false-positive cases are included among the identified cases. When the concern with false-negative cases is negligible (e.g., when the sensitivity is deemed sufficiently high so that the number of false-negative cases is minimal), a high PPV might be adequate to provide confidence in the validity of the outcome variable, whereas a moderate-to-low PPV might warrant complete verification of the outcome variable for all potential cases. When the extent

of false-positive cases and the extent of false-negative cases are of concern, sponsors should consider assessing all performance measures needed for quantitative bias analysis to evaluate the impact of outcome misclassification on the measure of association or take a more rigorous approach by validating the outcome variable for all potential cases and non-cases to accurately classify the outcome variable for each subject. Overall, the required extent of validation should be determined by necessary level of certainty and the implication of potential misclassification on study inference.

In general, sponsors should consider the trade-off between false-positive and false-negative cases when selecting an operational definition and identify the proper outcome validation approach to support internal validity. For example, to identify neural tube defects in infants, an operational definition that includes a spectrum of inpatient and outpatient diagnosis codes might have a high sensitivity, low specificity, and low PPV; restricting the operational definition to inpatient diagnosis codes only or a combination of diagnosis and procedure (e.g., surgical repair) codes might increase the PPV but miss a substantial proportion of true cases (low sensitivity). Because missing true cases is particularly a concern for infrequently reported outcomes, one approach is to select an operational definition of high sensitivity and perform complete verification of the outcome variable for all potential cases to maximize the likelihood that the true cases are all identified and that false-positive cases are minimized through validation. Unlike rare disease outcomes, when an outcome of interest involves a more common event (e.g., disease-specific hospitalization) or improvement or worsening of a condition, the operational definitions for common diagnoses are likely to generate false-positive and false-negative cases to a considerable extent because both true cases and true non-cases are prevalent. Therefore, it might be difficult to obtain accurate and complete information (e.g., laboratory test results, functional measures) for the operational definition to accurately classify cases and non-cases. For such outcomes, measuring PPV alone will be inadequate to inform outcome misclassification.

In scenarios where complete verification of the outcome variable for each study subject is infeasible, the performance of an operational outcome definition should be assessed in the proposed study population using a justified sampling strategy. As stated earlier, use of an operational definition that has been assessed in a prior study should ideally be in the same data source and in a similar study population, because the performance of an operational definition may vary substantially by data source and study scenario, and more recent data may be needed if there are secular trends in disease, diagnosis, and coding. The quality of prior studies used to establish sensitivity, specificity, and predictive values should be evaluated. In particular, the case definition used in the prior study to establish these measures should be compatible with the conceptual outcome definition developed for the proposed study. The applicability of these measures to the proposed study should be justified, and sensitivity analyses can be considered.

Without complete patient information and complete verification of the outcome variable, outcome misclassification remains a threat to the study internal validity, and the impact on the measure of association between exposure and outcome varies depending on whether the degree of misclassification differs between the exposure groups. Differential misclassification involves a complex interplay of differences in sensitivity, specificity, and disease prevalence between the exposure groups, and thus may bias the association either toward or away from the null. Because

it is difficult to predict the direction of the bias, differential misclassification is a concern for both safety and effectiveness studies. Unlike differential misclassification, non-differential misclassification tends to bias the association toward the null; as a result, a true risk might be missed in safety studies, whereas a larger study population might be needed to demonstrate the drug effect in effectiveness studies.

Non-differential outcome misclassification might occur when the outcome definition is not adequately refined and includes conditions that are not uniformly associated with the exposure of interest. For example, neural tube defects include primary neurulation defects and post-neurulation defects. Primary neurulation defects are directly attributed to failure of primary neurulation (i.e., neural tube closure), which occurs between approximately 18 and 28 days after fertilization. The pathophysiology of post-neurulation defects is less understood. Therefore, drug exposure during the critical period for primary neurulation in gestation might not affect post-neurulation in the same manner. When the outcome definition includes both primary and post-neurulation periods, the risk of primary neurulation defects, if any, is likely not detected.

Differential outcome misclassification might be minimized in studies in which the exposure status is blinded. However, even when data collection methods seem to preclude the likelihood of differential outcome misclassification, non-differential outcome misclassification is not guaranteed in the actual data of a particular study. For example, the physician who observed, diagnosed, and documented whether or not an outcome occurred could have been the same physician who made a decision as to which patients received the treatment meant to prevent that outcome, or the physician could have monitored disease progression or treatment side effects differently, given the knowledge as to which treatment they received. Biased misclassification can also result from public announcements of safety concerns with a particular drug if the data include events that occurred after the date of the public announcement. Therefore, the direction of the outcome misclassification bias might remain unpredictable when using real-world data. In addition, when more than one misclassification exists in a study, sponsors should consider how they might be related to each other. For example, whereas non-differential exposure misclassification and non-differential outcome misclassification each might bias the association toward the null, when the two misclassifications are dependent, overall it can create a bias away from the null (Lash et al. 2009). Therefore, when evaluating the implication of potential misclassification on study inference, sponsors should avoid overreliance on non-differential misclassification biasing toward the null. Under such circumstances, assessing the performance of the operational outcome definition according to exposure status in the proposed study population might be necessary.

Regarding outcome validation, sponsors should justify the proposed validation approach, such as validating the outcome variable for all potential cases or non-cases, versus assessing the performance of the proposed operational definition; if the latter will be done, justify what performance measures will be assessed. The protocol should include a detailed description of the outcome validation design, methods, and processes, as well as sampling strategy (if applicable). If a previously assessed operational definition is proposed, additional information should be provided, including: (1) data source and study population; (2) during what time frame validation was performed; (3) performance characteristics; (4) the reference standard against

which the performance was assessed; and (5) a discussion of whether prior validation data are applicable to the proposed study.

FDA recommends including a quantitative bias analysis in the protocol as a sensitivity analysis to demonstrate whether and how outcome misclassification might affect study results. The protocol should prespecify the indices (e.g., sensitivity, specificity, PPV, NPV) that will be used for quantitative bias analysis and describe how the selected indices will be measured in outcome validation.

4. Mortality as an Outcome

In the United States, death and cause of death are generally not included in electronic health care data, with exceptions being made for death occurring while a patient is under medical care. Ascertainment of death (fact of death and cause of death) can be accomplished through linkage with public or commercial vital statistics data sources, to increase the completeness and recency of the death variables. The use of external mortality data, however, is subject to all of the limitations of such data and data linkage methods (Haynes 2019; Navar et al. 2019; Curtis 2018). Careful documentation of mortality data quality and its implications should be included in the protocol.

If the death is not captured in the electronic health care data systems, patients who die after having been exposed to the study drug might be observed in electronic health care data as either not filing any further medical claims or not receiving any additional care past a particular date. For studies in which the outcome or outcomes of interest (e.g., myocardial infarction or stroke) include fatal outcomes, excluding patients who appear to be lost to follow-up at any time following their exposure to the study drug is likely to create bias. These patients should be included in searches of vital statistics systems to see whether their absence (disenrollment) from the system is because of death, and it may be necessary to classify their deaths as an outcome of interest in the absence of data to the contrary.

E. Covariate Ascertainment and Validation

For the purposes of this guidance, covariates in a particular study can include two types of elements: confounders and ***effect modifiers***.

1. Confounders

Information on potential confounders is collected in a nonrandomized study to support appropriate efforts to balance treatment and control groups in the analysis. Epidemiologic and statistical methods for identifying and handling confounding in studies will be addressed in future guidance documents on RWE study design.

After identifying the potential confounders in a study, the proposed data source should be evaluated to determine whether it is adequate to capture information on important factors which may contribute to confounding. These include confounders that are well-captured in the proposed data source (measured confounders) and those that are not well-captured (unmeasured

or imperfectly measured confounders). Examples of confounders that can be unmeasured or imperfectly measured in electronic health care data, especially in claims data, include race/ethnicity, family history of disease, lifestyle factors (e.g., smoking, alcohol use, nutrition intake, physical activity), certain physical measurements (e.g., body mass index), drugs obtained without insurance, and indication for drug use. FDA recommends considering potential linkages with other data sources or additional data collection to expand the capture of important confounders that are unmeasured or imperfectly measured in the original data source.

2. Effect Modifiers

Studies of drug effectiveness or safety usually report an average treatment effect, even though the same treatment can have different effects in different groups of people. Information on potential effect modifiers is used to better understand heterogeneity of treatment effect, the nonrandom, explainable variability in the direction and magnitude of treatment effects for individuals within a population (Velentgas et al. 2013). The potential for effect modification by demographic variables (e.g., age, gender, race, ethnicity) or pertinent comorbidities should be examined in the study, and relevant effect modifiers should be available in the chosen data source.

3. Validation of Confounders and Effect Modifiers

For all key covariates, including confounders and effect modifiers, FDA recommends providing and justifying the validity of operational definitions in the protocol and study report. If the measured covariates can change during a patient's follow-up period (time-varying covariates) and are important to the analysis, the protocol should describe whether and how frequently the information on time-varying covariates can be captured, particularly since capture of time-varying covariates in RWD can be differential by severity of illness (e.g., more testing in more seriously ill patients).

When evaluating the validity of covariate operational definitions, FDA recommends identifying the best reference data source based on the nature of the covariates. When validating operational definitions of covariates that are medical events or procedure utilizations (e.g., comorbidities, past medical history), the same principles apply as in Section V.D.3, Validation of Outcomes. For discussion on validating operational definitions of covariates that are associated with drug uses, such as concurrent medications or past drug uses, see Section V.C.5, Validation of Exposure. When assessing the validity of other covariate operational definitions, such as family history of disease, lifestyle factors, or indication for drug use, the appropriate reference may include a patient or provider survey or appropriate data linkages.

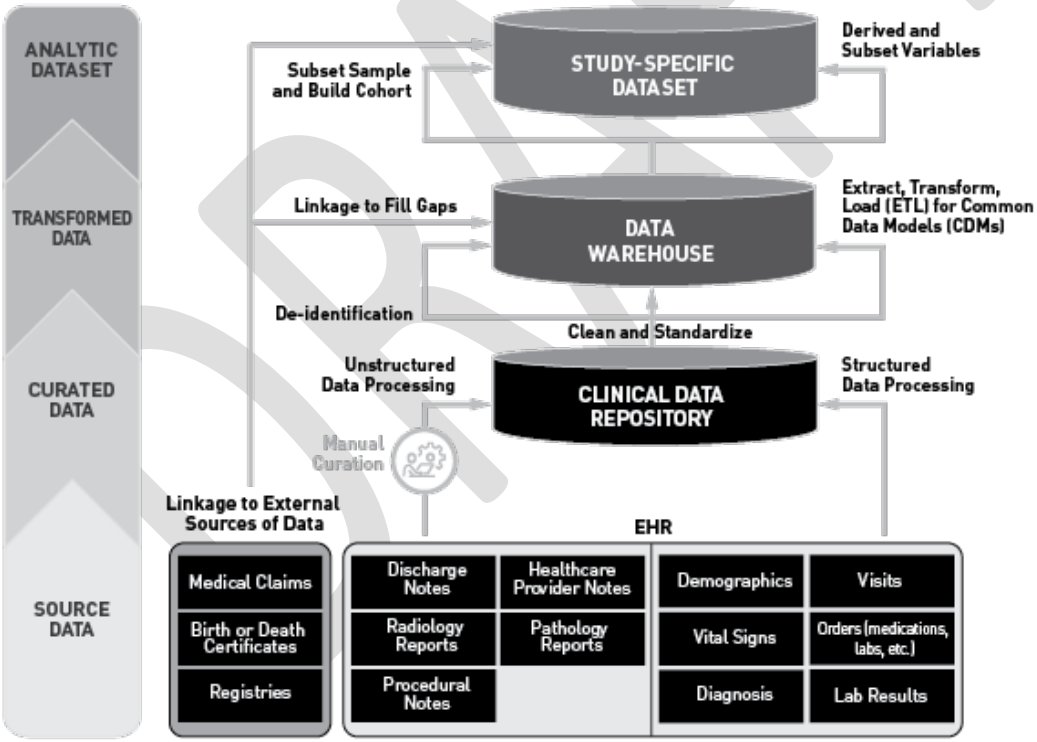
When supplemental information is needed to capture important covariates or is used for covariate validation, FDA recommends describing the likelihood of obtaining the supplemental information for the overall study population. If this supplemental information is only available for part of the study population, FDA recommends discussing the potential effect on internal validity in relevant study documents.

VI. DATA QUALITY DURING DATA ACCRUAL, CURATION, AND TRANSFORMATION INTO THE FINAL STUDY-SPECIFIC DATASET

This section discusses points for consideration when examining the quality of data over the course of the data life cycle. Although the data life cycle may vary depending on the type of data and setting (i.e., health care settings such as pharmacies, clinics, emergency departments and hospitals), in general, the life cycle involves multiple phases: data accrual from the original ***source data***; curation of data to the clinical ***data repository***; transformation and ***de-identification*** of data where necessary, creation of a ***data warehouse***; and production of a study-specific dataset for analysis (see Figure 1).

The concept of the data life cycle illustrates the iterative nature of the process for examining the quality of data. The process is not a one-time assessment; rather, it is an ongoing process in which data quality checks, cleansing[15], and monitoring occur at each phase in the cycle, and some checks may be repeated (i.e., occur in multiple phases of the cycle).

Figure 1: Illustrative Example of the Life Cycle of EHR Data[16]



[15] Data cleansing (sometimes referred to as data scrubbing) is the process of correcting or removing inaccurate data (or improperly formatted, duplicate data or records) from a database. The data requiring correction/removal is sometimes referred to as "dirty data." Data cleansing is an essential task for preserving data quality.

[16] This figure illustrates some of the processes applied to EHR data to produce a dataset that may be appropriate for research use (i.e., steps from original source data through the final analytic dataset). This figure shows processes for EHR data; the process may differ for claims data. Quality checks for each process step are described in this section.

Guidelines that evaluate the quality of EHRs and medical claims data primarily focus on distributed data networks in which disparate data sources are aggregated, linked, and processed to create a comprehensive data warehouse (Miksad and Abernethy 2018; Girman et al. 2018; Daniel et al. 2018; Kahn et al. 2016; Wang et al. 2017; Mahendraratnam et al. 2019). Although FDA does not endorse any particular set of guidelines or checklists, researchers should evaluate the completeness, accuracy, and ***plausibility*** of the data, including verifying data against its original source (e.g., discharge notes, pathology reports, registry records) and conforming to consensus-based data standards, where applicable. Researchers should provide scientific justifications for choosing these standards and should articulate how these standards are adequate to ensure the completeness, accuracy, and plausibility of the relevant data source.

The study protocol and analysis plan should specify the data provenance (curation and transformation procedures used throughout the data life cycle) and describe how these procedures could affect ***data integrity*** and the overall validity of the study. Below are points for consideration when examining data at each step in the data life cycle, including (A) characterizing the data with respect to completeness, ***conformance***, and plausibility of data values, (B) documenting the QA/QC plan that includes transformation processes; and (C) defining a set of procedures for ensuring data integrity.

A. Characterizing Data

The format and provenance of EHR and medical claims data can vary significantly across health care entities (e.g., insurer, practice, provider, data vendor). In general, researchers should address the procedures used to ensure completeness and accuracy of the data, as well as processes for data accrual, curation, and transformation over the data life cycle. The FDA recommends automated data quality reports that include the following characteristics and processes in a standardized way, when applicable to the chosen data source:

- Data accrual

  1. Methods for data retrieval and processes to minimize missing data extraction, implausible values, and data quality checks in data captured at the point of care (e.g., during clinical practice for manual or automated health care data collection processes) to ensure accuracy and completeness of core data elements.

  2. Provenance of core data elements to allow tracking of these elements back to their respective points of origin, with clear documentation of modifications that may have occurred.

  3. Timeliness of data availability, data years spanned, and continuity of coverage (e.g., median duration of patient enrollment).

  4. Handling data discrepancies and duplicate records. RWD may stem from multiple data streams, across various settings and platforms, which may present data discrepancies for the same variable (e.g., when the information for the same

element is entered differently in different data sources) or even duplicate records for the same patient within the same data source.

5. The reason for and timing of data error corrections implemented by data holders during the relevant period of data collection.

6. The reason for and timing of changes in processes implemented by data holders during the relevant period of data collection that may impact data accrual and/or data quality checks.

7. Any updates or changes in coding practices and versioning (e.g., International Classification of Diseases [ICD] diagnosis codes, Healthcare Common Procedure Coding System codes) across the study period that are relevant to variables of interest.

8. Any other changes in the data (e.g., collection, reporting, definitions) during the study period and their potential impact on the study results.

- Data curation

1. Routine migration of data from various sources over time.

2. Quality assurance (QA) testing and data quality checks employed across sites, as well as the criteria used in determining whether data quality techniques are appropriate for the intended purpose of the data.

3. Core data elements that are well-defined with consistent and known clinical meaning and understanding of data provenance, as well as documentation of clinical definitions used.

4. Assessment of completeness of data elements and trends over time.

5. Unstructured and structured data processing (e.g., abstraction and conversion of unstructured data to structured data), including manual versus automated techniques.

6. Harmonization of structured data across systems.

7. Conformance to open, consensus-based data curation standards, when applicable.

8. Accuracy of mappings (e.g., in the presence of different coding systems, such as Systematized Nomenclature of Medicine—Clinical Terms [SNOMED CT] versus ICD-10-CM).

9. Additional harmonization and mapping considerations, if applicable (if data spans multiple countries—e.g., U.K. data used in addition to U.S. data).

- Data transformation

    1. Implementation of the extract, transform, and load process applied to the whole repository population as part of data warehouse creation.

    2. De-identification of patient records and ability to re-identify unique patients in original source data without losing traceability.

    3. Algorithms used to transform and cleanse the data, as well as availability of standard operating procedures, including procedures for verifying the data.

    4. Data standardization (e.g., data types, sizes, formats) for internal consistency of data elements and semantics, including semantics of local codes to a target terminology (e.g., for laboratory data).

    5. When converting multiple data sources into a CDM, processes used for data transformation into a CDM (e.g., common terminology and structure), the comprehensiveness of the CDM (e.g., does the CDM contain the key data elements), approaches (e.g., algorithms/methods) for identification and handling of duplicate records within and across data sources, and potential impact of restricting to CDM on sample size and duration of patient follow-up or duration of drug exposure. See Section IV.B.3, Distributed Data Networks.

    6. Implementation of data checks pertaining to data model conformance errors.

    7. Data transformation processes used in preparation for data linkage. See Section IV.B.2, Data Linkage and Synthesis.

    8. Quality of record linkage (i.e., linking records from multiple datasets) and deduplication (i.e., finding duplicate records in a dataset) process, which may vary depending on the accuracy of the data used to perform the matches and the accuracy of the linkage algorithm.

    9. Quantification of errors (e.g., false matches, missed matches) that may lead to biased study findings. These are important when evaluating linkage quality (Harron et al. 2017). It is important to report details of the linkage algorithm and appropriate metrics (e.g., linkage error rates, match rates, comparison of characteristics of linked and unlinked data). Additional considerations include whether the error is random or nonrandom, potential bias, and impact on risk estimates and study findings.

    10. Procedures for adjudicating discrepancies in linked data as well as plans for handling linkage discrepancies (e.g., adjusting risk estimates for the linkage error).

- Study-specific analytic dataset

    1. Adherence to data specifications outlined in the study protocol and statistical analysis plan when compiling the analytic dataset.

    2. Additional study-specific data transformations, such as data transformations that are only done for a subset of patients of interest and that are not applied to all patient records in the data warehouse (e.g., manual extraction of data from unstructured textual pathology reports).

    3. Data checks implemented on the final analytic dataset for implausible values for data elements (e.g., height, weight, blood pressure), how such values are addressed, and the completeness of data for key analytic variables.

    4. The extent, percentage, and pattern of missingness and implausible data. Depending on the analysis plan's proposed method for handling missing data, imputations may be performed and included in the final analytic dataset and the type of imputation described.

B. Documentation of the QA/QC Plan

A QA/QC plan for construction of analytical data, the planned approach for handling quality control issues during analysis, and contemplation of differing levels of data quality by data element (and the potential implications on study findings) should be described in the study protocol and analysis plan. In general, activities to ensure the quality of the data before data-related activities are developed during the design of the study, and such activities, which include standardizing procedures for how to collect the data, may be regarded as QA (Szklo and Nieto 2006). Quality control consists of the decisions and steps taken from data collection through compilation of the final analytic dataset to ensure it meets prespecified standards and to ensure the processes used are reproducible. A multidisciplinary approach that includes clinical input is necessary to ensure adequate capture and handling of data, particularly for electronic health care systems, which inherently incorporate nuances and intricacies of health care delivery.

C. Documentation of Data Management Process

All manual and automated data retrieval and transformation processes should be thoroughly assessed from data collection through writing of the final study report to ensure data integrity. Researchers should ensure that curation and transformation processes do not alter the meaning of data or cause the loss of important contextual information. Descriptions of processes should include safeguards or checks to ensure that patient data are not duplicated or overrepresented. In addition, documentation of processes used to mine and evaluate unstructured data should describe the techniques employed (e.g., natural language processing) to abstract unstructured data (e.g., clinician notes) and supplement structured data (e.g., diagnostic codes).

Processes used for managing and preparing the final study-specific analytic dataset should be described in the study protocol or analysis plan. Analysts should have appropriate training or

experience with the data and software used to compile the analytic datasets. To facilitate FDA review, all submitted programs (e.g., those written by analysts) should be thoroughly annotated with comments that describe the intent or purpose of each data management and analysis step written in the program (e.g., annotate each data step in a statistical analysis program).

VII. GLOSSARY

Accuracy: Closeness of agreement between the measured value and the true value of what is intended to be measured.[17]

Artificial Intelligence (AI): The science and engineering of making intelligent machines, especially intelligent computer programs (McCarthy 2007).

Common Data Model (CDM): Standardizes a variety of electronic health care data sources into a common format to ensure interoperability across all sites providing data.[18]

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

Computable Phenotype: A clinical condition or characteristic that can be ascertained using a computerized query to an EHR system or clinical data repository (including disease registries, claims data) using a defined set of data elements and logical expressions. Computable phenotype definitions provide the specifications for identifying populations of patients with conditions of interest.[19]

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

Concomitant Medication: Prescription or nonprescription drugs or supplements used concurrently with the product of interest or comparator agent.

Conformance: "[D]ata congruence with standardized types, sizes, and formats" (Daniel et al. 2018).

Confounder (Confounding Factor): A variable that can be used to decrease confounding bias when properly adjusted for in an analysis. Confounding is the distortion of a measure of the effect of an exposure on an outcome because of the association of the exposure with other factors

[17] Adapted from the Joint Committee for Guides in Metrology guidance *International Vocabulary of Metrology—Basic and General Concepts and Associated Terms*, 3rd edition, 2012.

[18] Adapted from Sentinel System *Principles and Policies* (July 2019), available at https://www.sentinelinitiative.org/sites/default/files/About/Sentinel-System-Principles-and-Policies.pdf

[19] See the *NIH Collaboratory Living Textbook of Pragmatic Clinical Trials* chapter "Electronic Health Records-Based Phenotyping," available at https://rethinkingclinicaltrials.org/resources/ehr-phenotyping/.

that influence the occurrence of the outcome. Confounding occurs when all or part of the apparent association between the exposure and the outcome is in fact accounted for by other variables that affect the outcome and are not themselves affected by exposure (Porta 2014).

Continuity of Coverage: The period of time over which a patient is enrolled in a health care system and during which any medical service or drug prescription would be captured in that health care system's electronic record system.[20]

Covariate: A variable that is neither an exposure nor outcome of interest, but is measured to describe a population or because it may be a confounder or effect modifier to account for in study design or analysis.

Cumulative Dose: The total amount of the drug of interest (exposure) given to a patient over a specified period of time.[21]

Data Accrual: The process by which the data was collected.

Data Curation: Application of standards (e.g., Health Level 7, ICD-10-CM) to source data; for example, the application of codes to adverse events, disease staging, the progression of disease, and other medical and clinical concepts in an EHR.

Data Element: A piece of data corresponding to one patient within a data field (from Daniel, et al. 2018).

Data Integrity: The completeness, consistency, and accuracy of data.[22]

Data Repository: A database that consolidates data from disparate clinical sources, such as those within an EHR system, to provide a broader picture of the care a patient has received.[23]

Data Transformation: Includes data extraction, cleansing, and integration (e.g., into a CDM).

Data Warehouse: Consists of data from the data repository that has undergone data transformation and de-identification.

---

[20] See FDA guidance for industry *Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data* (May 2013).

[21] Adapted from the "NCI Dictionary of Cancer Terms," available at https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cumulative-dose.

[22] See FDA guidance for industry *Data Integrity and Compliance with Drug CGMP Questions and Answers* (December 2018).

[23] Adapted from Shortliffe, EH, and JJ Cimino, 2014, Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 4th Edition, New York (NY): Springer.

De-Identification: The process by which personal identifiers are removed from an individual's health information.[24]

Distributed Data Network: A network of multiple dispersed health care data sites providing the ability to query or analyze data from any or all sites.

Effect Modifier: A factor that biologically, clinically, socially, or otherwise alters the effects of another factor under study (Porta 2014).

Electronic Health Care Data: Analytic data that is an organized collection of automated health data available from computers or other electronic technological platforms.[25]

Electronic Health Record (EHR): An individual patient record contained within an EHR system. A typical individual EHR may include a patient's medical history, diagnoses, treatment plans, immunization dates, allergies, radiology images, pharmacy records, and laboratory and test results.[26]

Medical Claims Data: The compilation of information from medical claims that health care providers submit to insurers to receive payment for treatments and other interventions. Medical claims data use standardized medical codes, such as the World Health Organization's International Classification of Diseases Coding (ICD-CM) diagnosis codes, to identify diagnoses and treatments.[27]

Misclassification: The erroneous classification of an individual, value, or attribute into a category other than that to which it should be assigned (Porta 2014).

Missing Data: Data that would have been used in the study analysis but were not observed, collected, or accessible. This refers to information that is intended to be collected but is absent and information that is not intended to be collected and is therefore absent.

Negative Predictive Value (NPV): The probability that a subject does not have a disease when the classification result is negative.

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

---

[24] See Department of Health and Human Services *Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance With the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, available at https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

[25] Adapted from Hartzema, A, HH Tilson, and KA Chan, 2008, Pharmacoepidemiology and Therapeutic Risk Management, Cincinnati (OH): Harvey Whitney Books.

[26] See FDA guidance for industry *Use of Electronic Health Record Data in Clinical Investigations* (July 2018)

[27] See *Framework for FDA's Real-World Evidence Program* (December 2018)

Plausibility: The believability or truthfulness of data values (Kahn et al. 2016).

Positive Predictive Value (PPV): The probability that a subject has a disease when the classification result is positive.

Provenance: An audit trail that "accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place."[28]

Sensitivity: The probability that a classification result will be positive when the subject has the disease.

Source Data: All information in original records and certified copies of original records of clinical findings, observations, or other activities in a clinical study necessary for the reconstruction and evaluation of the study. Source data are contained in source documents (original records or certified copies).[29]

Specificity: The probability that a classification result will be negative when the subject does not have the disease.

Study Period: The calendar time range of data used for the study (Wang et al. 2017).

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.[30]

Validation: The process of establishing that a method is sound or that data are correctly measured, usually according to a reference standard.[31]

VIII. REFERENCES

Bustamante, R, A Earles, JD Murphy, AK Bryant, OV Patterson, AJ Gawron, T Kaltenbach, MA Whooley, DA Fisher, SD Saini, S Gupta, and L Liu, 2019, Ascertainment of Aspirin Exposure Using Structured and Unstructured Large-scale Electronic Health Record Data, Med Care, 57:e60–e64.

Carreras, G, M Simonetti, C Cricelli, and F Lapi, 2018, Deterministic and Probablistic Record Linkage: an Application to Primary Care Data, J Med Sys, 42(5):82.

[28] *Encyclopedia of Database Systems* definition of *data provenance*, available at https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_1305.

[29] See FDA guidance for industry *Use of Electronic Health Record Data in Clinical Investigations* (July 2018).

[30] See FDA technical specifications document *Study Data Technical Conformance Guide* (October 2019).

[31] Adapted from Porta, M, 2014, A Dictionary of Epidemiology, Sixth Edition, New York (NY): Oxford University Press.

Curtis, M, SD Griffith, M Tucker, MD Taylor, WB Capra, G Carrigan, B Holzman, AZ Torres, P You, B Arnieri, and AP Abernethy, 2018, Development and Validation of a High-Quality Composite Real-World Mortality Endpoint, Health Services Research, 53(6)Part I:4460-4476.

Daniel, G, C Silcox, J Bryan, M McClellan, M Romine, and K Frank, 2018, Characterizing RWD Quality and Relevancy for Regulatory Purposes, Duke Margolis Center for Health Policy, accessed January 9, 2019, https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf.

Girman, CJ, ME Ritchey, W Zhou, and NA Dreyer, 2019, Considerations in Characterizing Real-World Data Relevance and Quality for Regulatory Purposes: A Commentary, Pharmacoepidemiol Drug Saf, 28(4):439–442.

Harron, KL, JC Doidge, HE Knight, RE Gilbert, H Goldstein, DA Cromwell, and JH van der Meulen, 2017, A Guide to Evaluating Linkage Quality for the Analysis of Linked Data, Int J Epidemiol, 46(5):1699–1710.

Haynes, K, 2019, Mortality: The Final Outcome, Pharmacoepidemiol Drug Saf, epub ahead of print Jan 31, 2019, doi: 10.1002/pds.4715.

Kahn, MG, TJ Callahan, J Barnard, AE Bauck, J Brown, BN Davidson, H Estiri, C Goerg, E Holve, SG Johnson, ST Liaw, M Hamilton-Lopez, D Meeker, TC Ong, P Ryan, N Shang, NG Weiskopf, C Weng, MN Zozus, and L Schilling, 2016, A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, EGEMS, 4(1):1244.

Lash, TL, MP Fox, and AK Fink, 2009, Applying Quantitative Bias Analysis to Epidemiologic Data, New York (NY): Springer.

Mahendraratnam, N, C Silcox, K Mercon, A Kroetsch, M Romine, N Harrison, A Aten, R Sherman, G Daniel and M McClellan, 2019, Determining Real-World Data's Fitness for Use and the Role of Reliability, Duke Margolis Center for Health Policy, accessed July 24, 2020 https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf.

McCarthy, J, 2007, What Is Artificial Intelligence?, John McCarthy's Home Page, updated November 12, 2007, http://www-formal.stanford.edu/jmc/index.html.

Miksad, RA, and AP Abernethy, 2018, Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality, Clin Pharmacol Ther, 103(2):202–205.

National Institutes of Health Collaboratory, 2014, Assessing Data Quality for Healthcare Systems Data Used in Clinical Research, accessed August 27, 2019, https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf#search=Assessing%20data%20quality.

Navar, AM, ED Peterson, DL Steen, DM Wojdyla, RJ Sanchez, I Khan, X Song, ME Gold, and MJ Pencina, 2019, Evaluation of Mortality Data from the Social Security Administration Death Master File for Clinical Research, JAMA Cardiol, epub ahead of print Mar 6, 2019, doi: 10.1001/jamacardio.2019.0198.

Porta, M, 2014, A Dictionary of Epidemiology, Sixth Edition, New York (NY): Oxford University Press.

Richesson, RL, MM Smerek, and CC Blake, 2016, A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications, EGEMS, 4(3):1232.

Szklo, M, and FJ Nieto, 2006, Epidemiology: Beyond the Basics, 2nd Edition, Burlington (MA): Jones & Bartlett Learning.

Velentgas, P, NA Dreyer, P Nourjah, SR Smith, and MM Torchia, editors, 2013, Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide, AHRQ Publication No. 12(13)–EHC099, accessed January 9, 2019, https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/user-guide-observational-cer-130113.pdf.

Wang, SV, S Schneeweiss, ML Berger, J Brown, F de Vries, I Douglas, JJ Gagne, R Gini, O Klungel, CD Mullins, MD Nguyen, JA Rassen, L Smeeth, and M Sturkenboom, 2017, Reporting to Improve Reproducibility and Facilitate Validity Assessment in Healthcare Database Studies V1.0, Pharmacoepidemiol Drug Saf, 26(9):1018–1032.