
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DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.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be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance.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

I.	INTRODUCTION AND SCOPE	1
II.	BACKGROUND	3
III.	GENERAL CONSIDERATIONS	3
IV.	DATA SOURCES	4
A.	Relevance of Data Source.....	5
B.	Data Capture: General Discussion	5
1.	<i>Enrollment and Comprehensive Capture of Care.....</i>	<i>6</i>
2.	<i>Data Linkage and Synthesis.....</i>	<i>7</i>
3.	<i>Distributed Data Networks</i>	<i>7</i>
4.	<i>Computable Phenotypes.....</i>	<i>9</i>
5.	<i>Unstructured Data</i>	<i>9</i>
C.	Information Content and Missing Data: General Considerations	10
D.	Validation: General Considerations.....	10
V.	STUDY DESIGN ELEMENTS	13
A.	Definition of Time Periods	13
B.	Selection of Study Population	13
C.	Exposure Ascertainment and Validation.....	14
1.	<i>Definition of Exposure</i>	<i>14</i>
2.	<i>Ascertainment of Exposure: Data Source.....</i>	<i>14</i>
3.	<i>Ascertainment of Exposure: Duration</i>	<i>15</i>
4.	<i>Ascertainment of Exposure: Dose.....</i>	<i>16</i>
5.	<i>Validation of Exposure</i>	<i>16</i>
6.	<i>Dosing in Special Populations.....</i>	<i>17</i>
7.	<i>Other Considerations.....</i>	<i>17</i>
D.	Outcome Ascertainment and Validation.....	18
1.	<i>Definition of Outcomes of Interest.....</i>	<i>18</i>
2.	<i>Ascertainment of Outcomes</i>	<i>19</i>
3.	<i>Validation of Outcomes.....</i>	<i>20</i>
4.	<i>Mortality as an Outcome</i>	<i>23</i>
E.	Covariate Ascertainment and Validation	23
1.	<i>Confounders.....</i>	<i>23</i>
2.	<i>Effect Modifiers.....</i>	<i>24</i>
3.	<i>Validation of Confounders and Effect Modifiers</i>	<i>24</i>
VI.	DATA QUALITY DURING DATA ACCRUAL, CURATION, AND TRANSFORMATION INTO THE FINAL STUDY-SPECIFIC DATASET	25
A.	Characterizing Data.....	26

Contains Nonbinding Recommendations

Draft — Not for Implementation

B.	Documentation of the QA/QC Plan.....	29
C.	Documentation of Data Management Process.....	29
VII.	GLOSSARY.....	30
VIII.	REFERENCES.....	33

Contains Nonbinding Recommendations

Draft — Not for Implementation

1 **Real-World Data: Assessing Electronic Health Records and Medical**
2 **Claims Data To Support Regulatory Decision-Making for Drug and**
3 **Biological Products**
4 **Guidance for Industry¹**
5

6
7 This draft guidance, when finalized, will represent the current thinking of the Food and Drug
8 Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not
9 binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the
10 applicable statutes and regulations. To discuss an alternative approach, contact the FDA staff responsible
11 for this guidance as listed on the title page.
12

13
14
15 **I. INTRODUCTION AND SCOPE**
16

17 The 21st Century Cures Act (Cures Act),² signed into law on December 13, 2016, is intended to
18 accelerate medical product development and bring innovations faster and more efficiently to the
19 patients who need them. Among other provisions, the Cures Act added section 505F to the
20 Federal Food, Drug, and Cosmetic Act (FD&C Act) (21 U.S.C. 355g). Pursuant to this section,
21 FDA created a framework for a program to evaluate the potential use of real-world evidence
22 (RWE) to help support the approval of a new indication for a drug³ already approved under
23 section 505(c) of the FD&C Act or to help to support or satisfy postapproval study requirements
24 (RWE Program).⁴
25

26 FDA is issuing this guidance as part of its RWE Program and to satisfy, in part, the mandate
27 under section 505F of the FD&C Act to issue guidance about the use of RWE in regulatory
28 decision-making.⁵ The RWE Program will cover clinical studies that use real-world data (RWD)
29 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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

30 This guidance is intended to provide sponsors, researchers, and other interested stakeholders with
31 considerations when proposing to use **electronic health records**⁶ (EHRs) or **medical claims data**
32 in clinical studies⁷ to support a regulatory decision on effectiveness or safety.

33

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

35

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

38

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

41

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

50

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

53

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

56

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

59

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

62

63 This guidance does not provide recommendations on choice of study design or type of statistical
64 analysis, and it does not endorse any type of data source or study methodology. For all study
65 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>.

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

70
71 The contents of this document do not have the force and effect of law and are not meant to bind
72 the public in any way, unless specifically incorporated into a contract. This document is intended
73 only to provide clarity to the public regarding existing requirements under the law. FDA
74 guidance documents, including this guidance, should be viewed only as recommendations, unless
75 specific regulatory or statutory requirements are cited. The use of the word *should* in FDA
76 guidances means that something is suggested or recommended, but not required.

77
78

II. BACKGROUND

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

87
88 This guidance complements the 2013 guidance by expanding on certain aspects of that guidance
89 relating to the selection of data sources and also provides additional guidance for evaluating the
90 relevance and reliability of both EHRs and medical claims data for use in a clinical study. This
91 guidance also provides a broader overview of considerations relating to the use of EHR and
92 medical claims data in clinical studies more generally, including studies intended to inform
93 FDA's evaluation of product effectiveness.

III. GENERAL CONSIDERATIONS

94
95
96
97 For all studies using EHRs or medical claims data that will be submitted to FDA to support a
98 regulatory decision, sponsors should submit protocols and statistical analysis plans before
99 conducting the study. Sponsors seeking FDA input before conducting the study should request
100 comments or a meeting to discuss the study with the relevant FDA review division. All essential
101 elements of study design, analysis, conduct, and reporting should be predefined, and, for each
102 study element, the protocol and final study report should describe how that element was
103 ascertained from the selected RWD source, including applicable validation studies. More
104 information about study elements is provided in Section V, Study Design Elements.

105
106 This guidance provides recommendations on selecting data sources to maximize the
107 completeness and accuracy of the data derived from EHRs and medical claims for clinical
108 studies. The use of certain study design features or specific analyses to address misclassified or
109 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>.

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

- 113
- 114 1. The appropriateness and potential limitations of the data source for the study question
115 and to support key study elements.
- 116
- 117 2. Time periods for ascertainment of study design elements.
- 118
- 119 3. ***Conceptual definitions*** and ***operational definitions*** for study design elements (e.g.,
120 inclusion/exclusion criteria for study population, exposure, outcomes, covariates) and the
121 results of validation studies. See Section V, Study Design Elements, for examples of
122 conceptual and operational definitions for study design elements.
- 123
- 124 4. Quality assurance and quality control (QA/QC) procedures for data accrual, curation, and
125 transformation into the final study-specific dataset.
- 126
- 127

IV. DATA SOURCES

128
129
130 Protocols submitted to FDA should identify all data sources proposed for the study, as well as
131 other relevant descriptive information (discussed below). FDA does not endorse one data source
132 over another or seek to limit the possible sources of data that may be relevant to answering study
133 questions.

134
135 Each data source should be evaluated to determine whether the available information is
136 appropriate for addressing a specific study hypothesis. Because existing electronic health care
137 data were not developed to support regulatory submissions to FDA, it is important to understand
138 their potential limitations when they are used for that purpose. Examples of potential limitations
139 include:

- 140
- 141 1. The purpose of medical claims data is to support payment for care; claims may not
142 accurately reflect a particular disease, or a patient may have a particular disease or
143 condition that is not reflected in claims data.
- 144
- 145 2. EHR data are generated for use in clinical care and may also serve as a basis for billing
146 and for auditing of practice quality measures. Data recorded in an EHR system depend
147 on each health care system's practices for patient care and the clinical practices of its
148 providers. In addition, data collection is limited to the data captured within an EHR
149 system or network, and may not represent comprehensive care (e.g., care obtained outside
150 of the health care system).
- 151
- 152 3. For prospective clinical studies proposing to use EHRs, it may be possible to modify the
153 EHR system for the purpose of collecting additional patient data during routine care
154 through an add-on module to the EHR system. However, given the limited ability to add

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

157
158 The historical experience with and use of the selected data source for research purposes should
159 be described in the protocol. This description should include how well the selected data source
160 has been shown to capture study elements (e.g., inclusion and exclusion criteria, exposures,
161 outcomes, key covariates) and how the data can be validated for a particular research activity.

162

A. Relevance of the Data Source

164

165 There are differences in the practice of medicine around the world and between health care
166 systems that may affect the relevance of the data source to the study question. Patients in
167 different types of commercial or government health care payment programs can differ in a range
168 of characteristics, such as age, socioeconomic status, health conditions, risk factors, and other
169 potential **confounders**. Various factors in health care systems and insurance programs, such as
170 medication tiering (e.g., first-line, second-line), formulary decisions, and patient coverage, can
171 influence the degree to which patients on a given therapy in one health care system might differ
172 in disease severity, or other disease characteristics, from patients on the same therapy in another
173 health care system. It is also important to identify whether the data sources cover all populations
174 relevant to the study if those sources are to be used to examine the study hypothesis.

175

176 FDA recommends providing:

177

- 178 1. The reason for selecting the particular data sources to address the specific hypotheses.
- 179
180 2. Background information about the health care system, including (if available) any
181 specified method of diagnosis and preferred treatments for the disease of interest, and the
182 degree to which such information is collected and validated in the proposed data sources.
- 183
184 3. A description of prescribing and use practices in the health care system (if available),
185 including for approved indications, formulations, and doses.

186

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

189

B. Data Capture: General Discussion

191

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

200

Contains Nonbinding Recommendations

Draft — Not for Implementation

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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291

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

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

312
313 The latter type of networks, comprised of disparate data systems such as the Sentinel system, are
314 facilitated by the use of CDMs. Networks using CDMs also typically provide tools and
315 methodologies for analysis, a consistent level of data curation, and periodic revision of the data
316 model to incorporate new data concepts as needed by researchers. Additionally, methodologies
317 have been developed that allow the ability to translate data from one CDM to another, however
318 these involve additional data transformations, which present added quality considerations. Data
319 curation and transformation into a CDM, as well as general QA/QC processes, are discussed in
320 Section VI, Data Quality During Data Accrual, Curation, and Transformation into the Final
321 Study-Specific Dataset.

322
323 Distributed data networks are typically comprised of EHR, medical claim, or registry data.
324 However, combining many data sources, especially with the addition of data transformation into
325 a CDM, adds a layer of complexity that should be considered. Because there are many different
326 configurations of distributed health data networks, the configurations discussed in this guidance
327 should not be considered comprehensive.

328
329 Transforming disparate database structures into a common health network with a CDM allows
330 research across health care sites that would otherwise be more complex and costly. However,
331 CDMs can introduce additional challenges for researchers to consider. Many CDMs, including
332 those developed for FDA’s Sentinel system, Biologics Effectiveness and Safety Initiative, and
333 the National Patient-Centered Clinical Research Network, were created to satisfy a specific set of
334 research purposes; the choice of data captured in a CDM is optimized for the types of data
335 measures and detail needed for the intended use (e.g., Sentinel system for postmarket safety
336 surveillance to inform regulatory decision-making, the National Patient-Centered Clinical
337 Research Network for patient-centered outcomes research). Therefore, data in CDM-driven

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

344
345 Before using a CDM-driven network, data elements collected by the CDM should be
346 considered—including whether needed data elements exist in the model and, if so, whether they
347 are required or optional elements—to determine suitability for the study and whether identified
348 deficiencies can be addressed by supplementing with customized study-specific data elements,
349 collecting additional data, or using other data elements present in the dataset that are reasonable
350 proxies for the missing information. It should be noted, such workarounds would involve
351 additional considerations by the sponsor such as the work involved with validating proxy
352 endpoints or any human subject research considerations that involve additional data. Suitability
353 may also be improved with more flexible CDMs that are frequently expanded for new uses. For
354 information on proxy variables, see Section IV.C, Missing Data: General Considerations.

355 4. *Computable Phenotypes*

356
357
358 Standardized **computable phenotypes** can facilitate identification of similar patient populations
359 and enable efficient selection of populations for large-scale clinical studies across multiple health
360 care systems. A computable phenotype definition should include metadata and supporting
361 information about the definition, its intended use, the clinical rationale or research justification
362 for the definition, and data assessing validation in various health care settings (Richesson et al.
363 2016). The computable phenotype definition, composed of data elements and phenotype
364 algorithm, should be described in the protocol and study report and should also be available in a
365 computer-processable format. Clinical validation of the computable phenotype definition should
366 be described in the protocol and study report. For additional information on validation, see
367 Section IV.D, Validation: General Considerations.

368 5. *Unstructured Data*

369
370
371 Large amounts of key clinical data are unstructured data within EHRs, either as free text data
372 fields (such as physician notes) or as other non-standardized information in computer documents
373 (such as PDF-based radiology reports). To enhance the efficiency of data abstraction, a range of
374 approaches, including both existing and emerging technologies, are increasingly being used to
375 convert unstructured data into a computable format. More recent innovations include
376 technology-enabled abstraction whereby software provides a mechanism for human data
377 abstractors (e.g., tumor registrars) to do their work in a consistent and scalable fashion.

378
379 Technological advances in the field of **artificial intelligence** (AI) may permit more rapid
380 processing of unstructured electronic health care data. Advances include natural language
381 processing, machine learning, and particularly deep learning to: (1) extract data elements from
382 unstructured text in addition to structured fields in EHRs; (2) develop computer algorithms that

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

385
386 All of these methods are computer-assisted to various levels but currently require a significant
387 amount of human-aided curation and decision-making, injecting an additional level of data
388 variability and quality considerations into the final study-specific dataset. If the protocol
389 proposes to use AI or other derivation methods, the protocol should specify the assumptions and
390 parameters of the computer algorithms used, the data source from which the information was
391 used to build the algorithm, whether the algorithm was supervised (i.e., using input and review
392 by experts) or unsupervised, and the metrics associated with validation of the methods. Relevant
393 impacts on data quality should be documented in the protocol and analysis plan.

C. Missing Data: General Considerations

394
395
396
397 There are two broad cases in which information may be absent from the data sources. The first
398 case is when the information was intended to be collected (e.g., structured field present in the
399 EHR), but is absent from the data sources. This is an example of traditional *missing data*. The
400 second case is when the information was not intended to be collected in the EHR and medical
401 claims data and is therefore absent. It is important to distinguish between these two cases and
402 understand the reasons why information is present or absent in EHRs and medical claims. For
403 example, lack of information about the result of a laboratory test could be caused by different
404 circumstances: (1) the test might not have been ordered by the health care provider; (2) the test
405 might have been ordered but not conducted; (3) the test might have been performed, but the
406 result was not stored or captured in the data source; or (4) the test might have been performed
407 and the result was stored in the data source, but data were not in an accessible format, or lost in
408 the transformation and curation process when the final study-specific dataset was generated.
409 Because providers might order a laboratory test based on a patient's characteristics, the decision
410 not to order the test or a patient's decision to forgo the test may have implications on the data's
411 fitness for use in a proposed study.

412
413 As discussed above, data linkage is one way to address missing data. It may also be possible to
414 identify a variable that is a proxy for the missing data. An example of a potential proxy variable
415 includes low-income subsidy under the Medicare Part D prescription drug program as a proxy
416 for a patient's socioeconomic status.

417
418 The protocol and the statistical analysis plan should be developed and based on an understanding
419 of reasons for the presence and absence of information. Descriptive analyses should be included
420 to characterize the missing data. Assumptions regarding the missing data (e.g., missing at
421 random, missing not at random) underlying the statistical analysis for study end points and
422 important covariates should be supported and the implications of missing data considered.

D. Validation: General Considerations

423
424
425
426 Studies using EHR and medical claims data sources should include conceptual definitions for
427 important study variables, including study population inclusion and exclusion criteria, exposure,
428 outcome, and covariates. A conceptual definition should reflect current medical and scientific

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

432
433 An operational definition should be developed based on the conceptual definition to extract the
434 most complete and accurate data from the data source. In many studies using EHR or medical
435 claims data, the operational definition will be a code-based electronic algorithm using structured
436 data elements. In other studies, the operational definition may be derived from extracting
437 relevant information from unstructured data or constructing an algorithm that combines
438 structured and unstructured data elements. Operational definitions can also specify additional
439 data collection, such as a patient survey, when appropriate.

440
441 Because operational definitions are usually imperfect and cannot accurately classify the variable
442 of interest for every subject, a resulting *misclassification* can lead to false positives and false
443 negatives (Table 1) and may bias the association between exposure and outcome in a certain
444 direction and degree. Although complete verification¹⁰ of a variable of interest minimizes
445 misclassification and maximizes study internal validity, understanding the implications of
446 potential misclassification for study internal validity and study inference is the key step in
447 determining what variables of interest might require validation and to what extent. For example,
448 in a study to quantify a drug effect, internal validity should be optimized, and misclassification
449 of key variables should be minimized to accurately measure the association. Some
450 misclassification might be tolerable in some studies when the presence of misclassification is not
451 expected to change the interpretation of results (e.g., for signal detection, or when the
452 hypothesized effect size is large and the impact of misclassification on the measure of
453 association is deemed minimal).

454
455 To understand how potential misclassification of a variable of interest (e.g., exposure, outcome,
456 covariate) might impact the measure of association and the interpretation of results, sponsors
457 should consider: (1) the degree of misclassification; (2) differential versus non-differential
458 misclassification (e.g., differential misclassification of outcome by exposure); (3) dependent
459 versus independent misclassification (e.g., correlated misclassifications of exposure and outcome
460 when both are self-reported in the same survey); and (4) the direction toward which the
461 association between exposure and outcome might be biased.

462
463 Although complete verification of a study variable is considered the most rigorous approach,
464 there are scenarios where verifying a variable for every subject might not be feasible (e.g., a very
465 large study population, lack of reference standard¹¹ data for all study subjects) and assessing the
466 performance of the variable's operational definition might suffice. Based on the performance
467 measures described in Table 1, sponsors should consider whether validating the variable to a

¹⁰For 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.”

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

470
471 Because the performance of an operational definition is dependent on various factors, such as
472 data source, study population, study time frame, and choice of reference standard, FDA
473 recommends assessing the performance of operational definitions in an adequately large sample
474 of the study population as part of the proposed study, using justified sampling methods (e.g.,
475 random sampling, stratified sampling). If sponsors propose to use an operational definition that
476 has been assessed in a prior study, ideally those operational definitions assessed in the same data
477 source and in a similar study population should be selected. In addition, secular trends in
478 disease, diagnosis, and coding may necessitate assessment of the operational definition using
479 more recent data. The quality of prior studies used to establish *sensitivity*, *specificity*, and
480 predictive values should always be evaluated.

481
482 The protocol should include a detailed description of the planned validation, including
483 justification for the choice of a reference standard, validation approach, methods, processes, and
484 sampling strategy (if applicable). If a previously assessed operational definition is proposed,
485 additional information should be provided, including in what data source and study population
486 and during what time frame the assessment was conducted, the value of the assessed
487 performance measures, and a discussion of whether the performance measures are applicable to
488 the proposed study. FDA also recommends including in the protocol prespecified sensitivity
489 analyses to demonstrate whether and how bias, if present, might impact study findings based on
490 the validation data.

491
492 For further discussion about the validation of study design elements, see Section V.C.5,
493 Validation of Exposure; Section V.D.3, Validation of Outcomes; and Section V.E.3, Validation
494 of Confounders and Effect Modifiers.

495
496 **Table 1: Schematic Representation of the Calculation of Sensitivity, Specificity, Positive**
497 **Predictive Value (PPV), and Negative Predictive Value (NPV) for a Binary Variable**
498

Condition based on proposed operational definition	Condition based on reference standard		Total	
	Yes	No		
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 = $d/(b+d)$		

499

Contains Nonbinding Recommendations

Draft — Not for Implementation

500 **V. STUDY DESIGN ELEMENTS**

501
502 The ascertainment and validation of key study design elements are discussed in detail below.
503 The study questions of interest should be established first, and then the data source and study
504 design most appropriate for addressing these questions should be determined. The study should
505 not be designed to fit a specific data source, because the limitations of a specific data source may
506 restrict the options for study design and limit the inferences that can be drawn. Considerations
507 regarding study design and analysis when using RWD sources will be discussed in other RWE
508 guidance documents.

509 **A. Definition of Time Periods**

510
511
512 FDA recommends clearly defining the various time periods pertinent to the study design in the
513 protocol (e.g., time periods for identifying study population, defining inclusion and exclusion
514 criteria, assessing exposure, assessing outcome, assessing covariates, following up with patients).
515 The focus of the time scale (e.g., calendar time, age, time since exposure) should be explicitly
516 described with adequate detail on data availability of the time unit (e.g., year, month, day, hour,
517 minute) required to answer the study question.

518
519 The protocol should justify proposed time periods and the potential impact on study validity. For
520 example, justification should be provided regarding whether the time period before exposure is
521 appropriate for identifying the study population and the important baseline covariates, whether
522 the follow-up time is sufficient for observing the occurrence of study outcomes, and whether the
523 time period for updating information on time-dependent covariates is suitable to capture the
524 changes of those variables. In addition, when considering outcome definitions, disease onset
525 (e.g., early symptoms) may need to be distinguished from a confirmed diagnosis, as appropriate
526 to the study question. When defining the beginning and the end of the follow-up time for
527 outcome assessment, consider the biologically plausible time frame when the outcome, if
528 associated with the exposure, might be expected to occur.

529
530 The protocol should also address potential temporal changes in the standard of care, the
531 availability of other treatments, diagnosis criteria, and any other relevant factors that are
532 pertinent to the study question and design. Other relevant factors may include insurance
533 formulary changes (if known), step therapy, and laboratory assay changes. Before developing
534 the study approach, sponsors should discuss with the relevant FDA review division the capability
535 of data to capture such potential temporal changes and the impact of the potential temporal
536 changes on internal validity.

537 **B. Selection of Study Population**

538
539
540 The protocol should include a detailed description of methods for determining how inclusion and
541 exclusion criteria (e.g., demographic factors, medical condition, disease status, severity,
542 biomarkers) will be implemented to identify appropriate patients meeting these criteria from the
543 data source. The protocol should address the completeness and accuracy of the information
544 collected in the proposed data source to fulfill the inclusion and exclusion criteria.

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

551
552 In certain circumstances, key variables (e.g., gestational age for pregnancy studies) required to
553 fulfill the inclusion and exclusion criteria may be generated by the health care provider using
554 information available at the point of care. For example, health care providers may enter the
555 calculated gestational age in an EHR based on patient self-reported last menstrual period,
556 ultrasound dating, and other relevant information. If such data are used, the protocol should
557 describe the source of information and the methods health care providers use to generate the data
558 (if known).

C. Exposure Ascertainment and Validation

561
562 Considerations discussed in this section regarding exposure ascertainment in medical claims data
563 or EHRs primarily apply to noninterventional studies, given that the assignment of exposure is
564 documented in interventional studies.

1. Definition of Exposure

565
566 For the purposes of this guidance, the term *exposure* applies to the medical product or regimen of
567 interest being evaluated in the proposed study. The product of interest is referred to as *the*
568 *treatment*, and may be compared to no treatment, a placebo, standard of care, another treatment,
569 or a combination of the above. Other variables that could affect the study outcome are
570 considered covariates and are discussed in Section V.E, Covariate Ascertainment and Validation.
571 The exposure definition should include information about the drug dose, formulation, strength,
572 route, timing, frequency, and duration the product studied (if relevant). It may also be necessary
573 to describe the specific manufacturer of a product (e.g., when a proper name for a vaccine is used
574 by different manufacturers).

575
576 The description of exposure should include the intended or prescribed use of the product (e.g.,
577 the number, frequency, and specific doses), the period between initiation of exposure and the
578 earliest time one might reasonably expect to see an effect, and the expected duration of effect.
579 This will usually require an understanding of the pharmacological properties of the drug—for
580 example, that a one-time infusion to prevent osteoporosis may have an effect for several months.
581 See Section V.C.3, Ascertainment of Exposure: Duration, and Section V.C.4, Ascertainment of
582 Exposure: Dose.

2. Ascertainment of Exposure: Data Source

583
584 Sponsors should be able to demonstrate an ability to identify the specific products of interest in
585 the proposed data source, demonstrating that the data source contains data fields and codes that
586 allow identification of the specific products of interest (e.g., through specific coding). For
587
588
589
590

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

596
597 When relying on coded data, the operational exposure definitions should be based on the coding
598 system of the selected data source and reflect an understanding of the prescription, delivery, and
599 reimbursement characteristics of the drug (if applicable) in that data source. For example, in the
600 United States, the operational definition should include the appropriate pharmacy codes (NDC or
601 Healthcare Common Procedure Coding System) to capture the use of the drug in various
602 settings. This approach is particularly important in the case of non-oral drugs that may be
603 assigned different codes depending on how they are obtained. For example, patients using an
604 injectable drug can purchase it from the pharmacy, in which case the NDC code would be
605 recorded, or it can be administered by the provider for the patient and the drug and its
606 administration would be recorded using the HCPCS J code.¹²

607
608 It is also essential to report operational definitions and methods when combining information
609 from unstructured and structured data. Emerging methods may involve review of unstructured
610 information in medical records combined with pharmacy dispensing and physician prescribing
611 data and notes to provide an assessment of whether a person was prescribed and received the
612 medication of interest, as well as whether there are problems with the patient continuing the
613 medication. An example of such methods is found in ascertainment of aspirin exposure in a
614 retrospective cohort study of veterans undergoing usual care colonoscopy (Bustamente et al.
615 2019).

616
617 When using a medical claims data source, it is important to consider that there could be
618 dispensed prescriptions that were not associated with insurance claims if these uncaptured
619 prescriptions are relevant exposures for the study. Uncaptured prescriptions might include low-
620 cost generic drugs, drugs obtained through discount programs, samples provided by
621 pharmaceutical companies and dispensed by health care providers, and drugs sold via the internet
622 or patient out-of-pocket purchases. In addition, nonprescription drugs and dietary supplements
623 are not generally captured in electronic health care databases. It is important to address the
624 likelihood of incomplete exposure ascertainment and its effect on study validity, see Section
625 V.C.5, Validation of Exposure.

626 627 3. *Ascertainment of Exposure: Duration*

628
629 The data source should capture the relevant exposure duration (anticipated use of a product over
630 time). Given that some medical products are designed as one-time exposures (e.g., vaccines),
631 and other products may be intended for use over extended periods of time, the suitability of a
632 data source will vary with the specific medical product under investigation. FDA recommends
633 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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

644
645 Because patients may not refill their prescriptions exactly on time or, alternatively, may refill
646 their prescriptions early, gaps or stockpiling in therapy may exist and may be reflected in the
647 data.¹³ FDA recommends describing and justifying in the protocol how researchers will measure
648 use, address potential gaps in therapy in the data source, and handle refill stockpiling if there are
649 early refills. Intermittent therapies (e.g., drugs used to treat pain on an as-needed basis) and
650 therapies for which samples are often provided to patients (e.g., expensive drugs, drugs that are
651 new to the market) present challenges in accurately assessing the actual exposure and duration of
652 exposure, see Section V.C.5, Validation of Exposure.

653 654 4. *Ascertainment of Exposure: Dose*

655
656 Data about exposure should include information about dose. Depending on the exposure and the
657 question of interest in the study, it may be useful to describe the dose of each administration or a
658 daily dose, as well as an estimated *cumulative dose*.

659
660 It is reasonable to begin with the dose information provided in the data source, and then discuss
661 in the protocol or study report the specific assumptions made when estimating the dose of the
662 exposures of interest, especially for pediatric patients. See Section V.C.6, Dosing in Special
663 Populations. It is also important to report how different dosage forms (e.g., parenteral versus
664 oral) will figure into the dose calculation if multiple forms are available.

665 666 5. *Validation of Exposure*

667
668 Other than for medications administered in hospital settings or infusion settings, electronic health
669 care data capture prescriptions of drugs and the dispensing of drugs to patients, but generally do
670 not capture actual patient drug exposure because this depends on patients obtaining and using the
671 prescribed therapy.

672
673 Validation ideally involves a comparison of the exposure classification in the proposed data
674 source with a reference data source,¹⁴ and produces estimates of misclassification that can be
675 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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

691
692 FDA recommends documenting the methods used to calculate and validate duration, dose,
693 switching, and other characteristics of exposure. Validation and misclassification issues should
694 be addressed in appropriate study documents.

695

696 6. *Dosing in Special Populations*

697

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

708

709 7. *Other Considerations*

710

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

719

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

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

726

727

D. Outcome Ascertainment and Validation

728

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

744

745

1. Definition of Outcomes of Interest

746

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

759

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

Contains Nonbinding Recommendations

Draft — Not for Implementation

768 criteria that are likely to be consistently captured in RWD, such as stroke, myocardial infarction
769 or pulmonary embolism, compared to outcomes that are more subjective or scaled in nature, such
770 as worsening of joint pain in rheumatoid arthritis or worsening of depression symptoms.
771 Sponsors should discuss proposed outcomes definitions with the FDA review division.
772

2. *Ascertainment of Outcomes*

773
774
775 To help identify potential cases in the selected data source and study population, operational
776 definitions using diagnosis and procedure codes (e.g., ICD-9-CM, ICD-10), laboratory tests (e.g.,
777 LOINC) and values, or unstructured data (e.g., physician's encounter notes, radiology and
778 pathology reports) should be developed based on the conceptual definition of the outcome of
779 interest. If the operational definition includes information abstracted from unstructured data in
780 the EHR or another data source (e.g., mention of spina bifida in birth certificate records for the
781 identification of neural tube defects in infants), the protocol should provide a detailed description
782 and rationale for the methods and tools used to process the unstructured data and the validation
783 of those methods. See Section IV.B.5, Unstructured Data, for additional information on
784 unstructured data. When patient- or physician-generated data (e.g., data required for subjective
785 end points) are proposed to assess the outcome of interest or to complement operational
786 definitions, the protocol should specify how the outcome measure (e.g., sign score, severity
787 index) will be defined and constructed and validated, if applicable, and how the data will be
788 collected.
789

790 The sensitivity and specificity of an operational definition are imperfect when there is outcome
791 misclassification. Given that it is usually not possible for sensitivity and specificity to be perfect
792 (i.e., 100%), outcome misclassification might result in both false positives and false negatives.
793 FDA recommends considering the potential impact of outcome misclassification on study
794 validity when developing or selecting an operational definition for the proposed study. For
795 example, when studying infrequently occurring outcomes in a cohort study, given the low
796 prevalence of the outcome event, it is important to achieve high specificity to minimize false-
797 positive cases and high sensitivity so that more true cases can be captured.
798

799 Operational definitions developed for one data source or study population might not perform as
800 well in other sources or populations, due to database-specific sensitivity and specificity as well
801 as variable disease prevalence. **Positive predictive value** (PPV) and **negative predictive value**
802 (NPV) are related to sensitivity and specificity and are a direct function of prevalence of the
803 outcome in the population in which the predictive values are measured. Therefore, PPV and
804 NPV are variable by data source and study population characteristics (e.g., demographic factors,
805 underlying diseases, comorbidities, clinical settings).
806

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

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

817

818 3. *Validation of Outcomes*

819

820 FDA expects validation of the outcome variable to minimize outcome misclassification.

821 Although complete verification of the outcome variable is considered the most rigorous
822 approach, there are scenarios where verifying outcome for every subject might not be feasible
823 and assessing the performance of the operational definition of the outcome might suffice.

824 Outcome validation involves using a clinically appropriate conceptual outcome definition to
825 determine whether a patient's status, classified by an operational definition, truly represents the
826 outcome of interest, typically by reviewing clinical details recorded in the patient's medical
827 records in either electronic or paper format.

828

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

844

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

852

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

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

867
868 In general, sponsors should consider the trade-off between false-positive and false-negative
869 cases when selecting an operational definition and identify the proper outcome validation
870 approach to support internal validity. For example, to identify neural tube defects in infants, an
871 operational definition that includes a spectrum of inpatient and outpatient diagnosis codes
872 might have a high sensitivity, low specificity, and low PPV; restricting the operational
873 definition to inpatient diagnosis codes only or a combination of diagnosis and procedure (e.g.,
874 surgical repair) codes might increase the PPV but miss a substantial proportion of true cases
875 (low sensitivity). Because missing true cases is particularly a concern for infrequently reported
876 outcomes, one approach is to select an operational definition of high sensitivity and perform
877 complete verification of the outcome variable for all potential cases to maximize the likelihood
878 that the true cases are all identified and that false-positive cases are minimized through
879 validation. Unlike rare disease outcomes, when an outcome of interest involves a more
880 common event (e.g., disease-specific hospitalization) or improvement or worsening of a
881 condition, the operational definitions for common diagnoses are likely to generate false-
882 positive and false-negative cases to a considerable extent because both true cases and true non-
883 cases are prevalent. Therefore, it might be difficult to obtain accurate and complete
884 information (e.g., laboratory test results, functional measures) for the operational definition to
885 accurately classify cases and non-cases. For such outcomes, measuring PPV alone will be
886 inadequate to inform outcome misclassification.

887
888 In scenarios where complete verification of the outcome variable for each study subject is
889 infeasible, the performance of an operational outcome definition should be assessed in the
890 proposed study population using a justified sampling strategy. As stated earlier, use of an
891 operational definition that has been assessed in a prior study should ideally be in the same data
892 source and in a similar study population, because the performance of an operational definition
893 may vary substantially by data source and study scenario, and more recent data may be needed
894 if there are secular trends in disease, diagnosis, and coding. The quality of prior studies used to
895 establish sensitivity, specificity, and predictive values should be evaluated. In particular, the
896 case definition used in the prior study to establish these measures should be compatible with
897 the conceptual outcome definition developed for the proposed study. The applicability of these
898 measures to the proposed study should be justified, and sensitivity analyses can be considered.

899
900 Without complete patient information and complete verification of the outcome variable,
901 outcome misclassification remains a threat to the study internal validity, and the impact on the
902 measure of association between exposure and outcome varies depending on whether the degree
903 of misclassification differs between the exposure groups. Differential misclassification involves
904 a complex interplay of differences in sensitivity, specificity, and disease prevalence between the
905 exposure groups, and thus may bias the association either toward or away from the null. Because

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

911
912 Non-differential outcome misclassification might occur when the outcome definition is not
913 adequately refined and includes conditions that are not uniformly associated with the exposure of
914 interest. For example, neural tube defects include primary neurulation defects and post-
915 neurulation defects. Primary neurulation defects are directly attributed to failure of primary
916 neurulation (i.e., neural tube closure), which occurs between approximately 18 and 28 days after
917 fertilization. The pathophysiology of post-neurulation defects is less understood. Therefore,
918 drug exposure during the critical period for primary neurulation in gestation might not affect
919 post-neurulation in the same manner. When the outcome definition includes both primary and
920 post-neurulation periods, the risk of primary neurulation defects, if any, is likely not detected.

921
922 Differential outcome misclassification might be minimized in studies in which the exposure
923 status is blinded. However, even when data collection methods seem to preclude the likelihood
924 of differential outcome misclassification, non-differential outcome misclassification is not
925 guaranteed in the actual data of a particular study. For example, the physician who observed,
926 diagnosed, and documented whether or not an outcome occurred could have been the same
927 physician who made a decision as to which patients received the treatment meant to prevent that
928 outcome, or the physician could have monitored disease progression or treatment side effects
929 differently, given the knowledge as to which treatment they received. Biased misclassification
930 can also result from public announcements of safety concerns with a particular drug if the data
931 include events that occurred after the date of the public announcement. Therefore, the direction
932 of the outcome misclassification bias might remain unpredictable when using real-world data. In
933 addition, when more than one misclassification exists in a study, sponsors should consider how
934 they might be related to each other. For example, whereas non-differential exposure
935 misclassification and non-differential outcome misclassification each might bias the association
936 toward the null, when the two misclassifications are dependent, overall it can create a bias away
937 from the null (Lash et al. 2009). Therefore, when evaluating the implication of potential
938 misclassification on study inference, sponsors should avoid overreliance on non-differential
939 misclassification biasing toward the null. Under such circumstances, assessing the performance
940 of the operational outcome definition according to exposure status in the proposed study
941 population might be necessary.

942
943 Regarding outcome validation, sponsors should justify the proposed validation approach, such as
944 validating the outcome variable for all potential cases or non-cases, versus assessing the
945 performance of the proposed operational definition; if the latter will be done, justify what
946 performance measures will be assessed. The protocol should include a detailed description of
947 the outcome validation design, methods, and processes, as well as sampling strategy (if
948 applicable). If a previously assessed operational definition is proposed, additional information
949 should be provided, including: (1) data source and study population; (2) during what time frame
950 validation was performed; (3) performance characteristics; (4) the reference standard against

Contains Nonbinding Recommendations

Draft — Not for Implementation

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

953
954 FDA recommends including a quantitative bias analysis in the protocol as a sensitivity analysis
955 to demonstrate whether and how outcome misclassification might affect study results. The
956 protocol should prespecify the indices (e.g., sensitivity, specificity, PPV, NPV) that will be used
957 for quantitative bias analysis and describe how the selected indices will be measured in outcome
958 validation.

959 960 **4. *Mortality as an Outcome***

961
962 In the United States, death and cause of death are generally not included in electronic health care
963 data, with exceptions being made for death occurring while a patient is under medical care.
964 Ascertainment of death (fact of death and cause of death) can be accomplished through linkage
965 with public or commercial vital statistics data sources, to increase the completeness and recency
966 of the death variables. The use of external mortality data, however, is subject to all of the
967 limitations of such data and data linkage methods (Haynes 2019; Navar et al. 2019; Curtis 2018).
968 Careful documentation of mortality data quality and its implications should be included in the
969 protocol.

970
971 If the death is not captured in the electronic health care data systems, patients who die after
972 having been exposed to the study drug might be observed in electronic health care data as either
973 not filing any further medical claims or not receiving any additional care past a particular date.
974 For studies in which the outcome or outcomes of interest (e.g., myocardial infarction or stroke)
975 include fatal outcomes, excluding patients who appear to be lost to follow-up at any time
976 following their exposure to the study drug is likely to create bias. These patients should be
977 included in searches of vital statistics systems to see whether their absence (disenrollment) from
978 the system is because of death, and it may be necessary to classify their deaths as an outcome of
979 interest in the absence of data to the contrary.

980 981 **E. *Covariate Ascertainment and Validation***

982
983 For the purposes of this guidance, covariates in a particular study can include two types of
984 elements: confounders and *effect modifiers*.

985 986 **1. *Confounders***

987
988 Information on potential confounders is collected in a nonrandomized study to support
989 appropriate efforts to balance treatment and control groups in the analysis. Epidemiologic and
990 statistical methods for identifying and handling confounding in studies will be addressed in
991 future guidance documents on RWE study design.

992
993 After identifying the potential confounders in a study, the proposed data source should be
994 evaluated to determine whether it is adequate to capture information on important factors which
995 may contribute to confounding. These include confounders that are well-captured in the
996 proposed data source (measured confounders) and those that are not well-captured (unmeasured

Contains Nonbinding Recommendations

Draft — Not for Implementation

997 or imperfectly measured confounders). Examples of confounders that can be unmeasured or
998 imperfectly measured in electronic health care data, especially in claims data, include
999 race/ethnicity, family history of disease, lifestyle factors (e.g., smoking, alcohol use, nutrition
1000 intake, physical activity), certain physical measurements (e.g., body mass index), drugs obtained
1001 without insurance, and indication for drug use. FDA recommends considering potential linkages
1002 with other data sources or additional data collection to expand the capture of important
1003 confounders that are unmeasured or imperfectly measured in the original data source.

1004

1005 2. *Effect Modifiers*

1006

1007 Studies of drug effectiveness or safety usually report an average treatment effect, even though
1008 the same treatment can have different effects in different groups of people. Information on
1009 potential effect modifiers is used to better understand heterogeneity of treatment effect, the
1010 nonrandom, explainable variability in the direction and magnitude of treatment effects for
1011 individuals within a population (Velentgas et al. 2013). The potential for effect modification by
1012 demographic variables (e.g., age, gender, race, ethnicity) or pertinent comorbidities should be
1013 examined in the study, and relevant effect modifiers should be available in the chosen data
1014 source.

1015

1016 3. *Validation of Confounders and Effect Modifiers*

1017

1018 For all key covariates, including confounders and effect modifiers, FDA recommends providing
1019 and justifying the validity of operational definitions in the protocol and study report. If the
1020 measured covariates can change during a patient's follow-up period (time-varying covariates)
1021 and are important to the analysis, the protocol should describe whether and how frequently the
1022 information on time-varying covariates can be captured, particularly since capture of time-
1023 varying covariates in RWD can be differential by severity of illness (e.g., more testing in more
1024 seriously ill patients).

1025

1026 When evaluating the validity of covariate operational definitions, FDA recommends identifying
1027 the best reference data source based on the nature of the covariates. When validating operational
1028 definitions of covariates that are medical events or procedure utilizations (e.g., comorbidities,
1029 past medical history), the same principles apply as in Section V.D.3, Validation of Outcomes.
1030 For discussion on validating operational definitions of covariates that are associated with drug
1031 uses, such as concurrent medications or past drug uses, see Section V.C.5, Validation of
1032 Exposure. When assessing the validity of other covariate operational definitions, such as family
1033 history of disease, lifestyle factors, or indication for drug use, the appropriate reference may
1034 include a patient or provider survey or appropriate data linkages.

1035

1036 When supplemental information is needed to capture important covariates or is used for
1037 covariate validation, FDA recommends describing the likelihood of obtaining the supplemental
1038 information for the overall study population. If this supplemental information is only available
1039 for part of the study population, FDA recommends discussing the potential effect on internal
1040 validity in relevant study documents.

1041

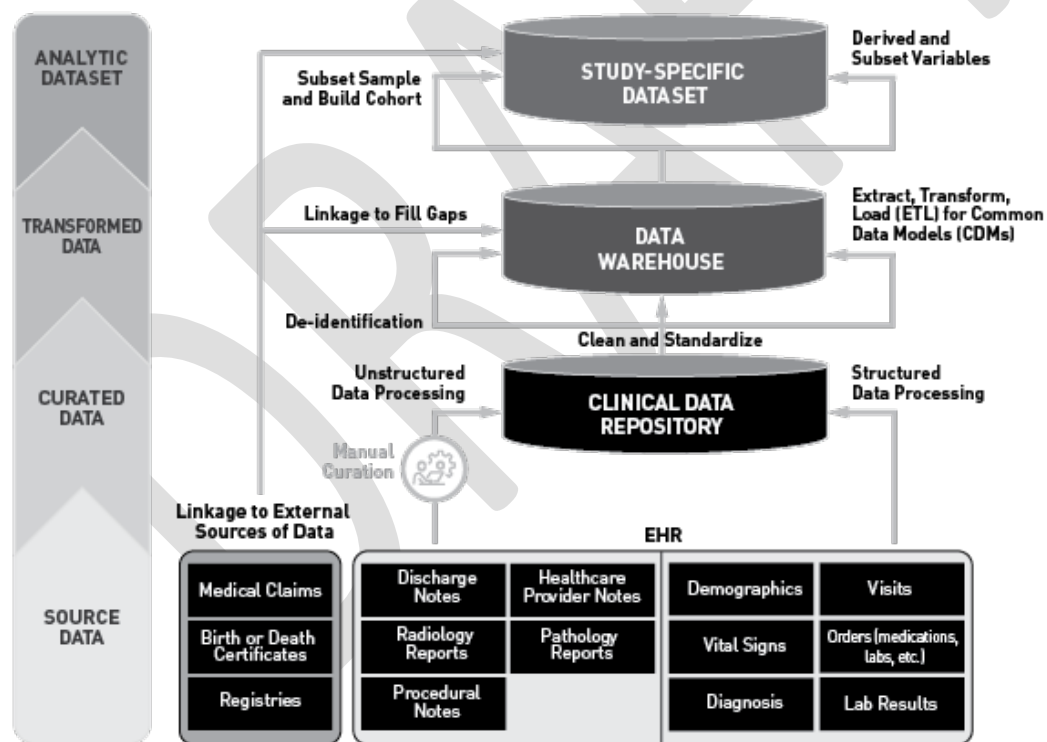
1042

1043 **VI. DATA QUALITY DURING DATA ACCRUAL, CURATION, AND**
 1044 **TRANSFORMATION INTO THE FINAL STUDY-SPECIFIC DATASET**
 1045

1046 This section discusses points for consideration when examining the quality of data over the
 1047 course of the data life cycle. Although the data life cycle may vary depending on the type of data
 1048 and setting (i.e., health care settings such as pharmacies, clinics, emergency departments and
 1049 hospitals), in general, the life cycle involves multiple phases: data accrual from the original
 1050 *source data*; curation of data to the clinical *data repository*; transformation and *de-identification*
 1051 of data where necessary, creation of a *data warehouse*; and production of a study-specific
 1052 dataset for analysis (see Figure 1).
 1053

1054 The concept of the data life cycle illustrates the iterative nature of the process for examining the
 1055 quality of data. The process is not a one-time assessment; rather, it is an ongoing process in
 1056 which data quality checks, cleansing¹⁵, and monitoring occur at each phase in the cycle, and
 1057 some checks may be repeated (i.e., occur in multiple phases of the cycle).
 1058

1059 **Figure 1: Illustrative Example of the Life Cycle of EHR Data¹⁶**
 1060



1061

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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

1062 Guidelines that evaluate the quality of EHRs and medical claims data primarily focus on
1063 distributed data networks in which disparate data sources are aggregated, linked, and processed
1064 to create a comprehensive data warehouse (Miksad and Abernethy 2018; Girman et al. 2018;
1065 Daniel et al. 2018; Kahn et al. 2016; Wang et al. 2017; Mahendraratnam et al. 2019). Although
1066 FDA does not endorse any particular set of guidelines or checklists, researchers should evaluate
1067 the completeness, accuracy, and *plausibility* of the data, including verifying data against its
1068 original source (e.g., discharge notes, pathology reports, registry records) and conforming to
1069 consensus-based data standards, where applicable. Researchers should provide scientific
1070 justifications for choosing these standards and should articulate how these standards are adequate
1071 to ensure the completeness, accuracy, and plausibility of the relevant data source.
1072

1073 The study protocol and analysis plan should specify the data provenance (curation and
1074 transformation procedures used throughout the data life cycle) and describe how these
1075 procedures could affect *data integrity* and the overall validity of the study. Below are points for
1076 consideration when examining data at each step in the data life cycle, including (A)
1077 characterizing the data with respect to completeness, *conformance*, and plausibility of data
1078 values, (B) documenting the QA/QC plan that includes transformation processes; and (C)
1079 defining a set of procedures for ensuring data integrity.
1080

A. Characterizing Data

1081
1082
1083 The format and provenance of EHR and medical claims data can vary significantly across health
1084 care entities (e.g., insurer, practice, provider, data vendor). In general, researchers should
1085 address the procedures used to ensure completeness and accuracy of the data, as well as
1086 processes for data accrual, curation, and transformation over the data life cycle. The FDA
1087 recommends automated data quality reports that include the following characteristics and
1088 processes in a standardized way, when applicable to the chosen data source:
1089

- 1090 • Data accrual
- 1091
- 1092 1. Methods for data retrieval and processes to minimize missing data extraction,
1093 implausible values, and data quality checks in data captured at the point of care
1094 (e.g., during clinical practice for manual or automated health care data collection
1095 processes) to ensure accuracy and completeness of core data elements.
1096
- 1097 2. Provenance of core data elements to allow tracking of these elements back to their
1098 respective points of origin, with clear documentation of modifications that may
1099 have occurred.
1100
- 1101 3. Timeliness of data availability, data years spanned, and continuity of coverage
1102 (e.g., median duration of patient enrollment).
1103
- 1104 4. Handling data discrepancies and duplicate records. RWD may stem from
1105 multiple data streams, across various settings and platforms, which may present
1106 data discrepancies for the same variable (e.g., when the information for the same

Contains Nonbinding Recommendations

Draft — Not for Implementation

- 1107 element is entered differently in different data sources) or even duplicate records
1108 for the same patient within the same data source.
1109
- 1110 5. The reason for and timing of data error corrections implemented by data holders
1111 during the relevant period of data collection.
1112
 - 1113 6. The reason for and timing of changes in processes implemented by data holders
1114 during the relevant period of data collection that may impact data accrual and/or
1115 data quality checks.
1116
 - 1117 7. Any updates or changes in coding practices and versioning (e.g., International
1118 Classification of Diseases [ICD] diagnosis codes, Healthcare Common Procedure
1119 Coding System codes) across the study period that are relevant to variables of
1120 interest.
1121
 - 1122 8. Any other changes in the data (e.g., collection, reporting, definitions) during the
1123 study period and their potential impact on the study results.
1124
- 1125 • Data curation
- 1126 1. Routine migration of data from various sources over time.
1127
 - 1128 2. Quality assurance (QA) testing and data quality checks employed across sites, as
1129 well as the criteria used in determining whether data quality techniques are
1130 appropriate for the intended purpose of the data.
1131
 - 1132 3. Core data elements that are well-defined with consistent and known clinical
1133 meaning and understanding of data provenance, as well as documentation of
1134 clinical definitions used.
1135
 - 1136 4. Assessment of completeness of data elements and trends over time.
1137
 - 1138 5. Unstructured and structured data processing (e.g., abstraction and conversion of
1139 unstructured data to structured data), including manual versus automated
1140 techniques.
1141
 - 1142 6. Harmonization of structured data across systems.
1143
 - 1144 7. Conformance to open, consensus-based data curation standards, when applicable.
1145
 - 1146 8. Accuracy of mappings (e.g., in the presence of different coding systems, such as
1147 Systematized Nomenclature of Medicine—Clinical Terms [SNOMED CT] versus
1148 ICD-10-CM).
1149
 - 1150 9. Additional harmonization and mapping considerations, if applicable (if data spans
1151 multiple countries—e.g., U.K. data used in addition to U.S. data).
1152

Contains Nonbinding Recommendations

Draft — Not for Implementation

1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198

- 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).

Contains Nonbinding Recommendations

Draft — Not for Implementation

- 1199 • Study-specific analytic dataset
1200
1201 1. Adherence to data specifications outlined in the study protocol and statistical
1202 analysis plan when compiling the analytic dataset.
1203
1204 2. Additional study-specific data transformations, such as data transformations that
1205 are only done for a subset of patients of interest and that are not applied to all
1206 patient records in the data warehouse (e.g., manual extraction of data from
1207 unstructured textual pathology reports).
1208
1209 3. Data checks implemented on the final analytic dataset for implausible values for
1210 data elements (e.g., height, weight, blood pressure), how such values are
1211 addressed, and the completeness of data for key analytic variables.
1212
1213 4. The extent, percentage, and pattern of missingness and implausible data.
1214 Depending on the analysis plan’s proposed method for handling missing data,
1215 imputations may be performed and included in the final analytic dataset and the
1216 type of imputation described.
1217

B. Documentation of the QA/QC Plan

1218
1219
1220 A QA/QC plan for construction of analytical data, the planned approach for handling quality
1221 control issues during analysis, and contemplation of differing levels of data quality by data
1222 element (and the potential implications on study findings) should be described in the study
1223 protocol and analysis plan. In general, activities to ensure the quality of the data before data-
1224 related activities are developed during the design of the study, and such activities, which include
1225 standardizing procedures for how to collect the data, may be regarded as QA (Szklo and Nieto
1226 2006). Quality control consists of the decisions and steps taken from data collection through
1227 compilation of the final analytic dataset to ensure it meets prespecified standards and to ensure
1228 the processes used are reproducible. A multidisciplinary approach that includes clinical input is
1229 necessary to ensure adequate capture and handling of data, particularly for electronic health care
1230 systems, which inherently incorporate nuances and intricacies of health care delivery.
1231

C. Documentation of Data Management Process

1232
1233
1234 All manual and automated data retrieval and transformation processes should be thoroughly
1235 assessed from data collection through writing of the final study report to ensure data integrity.
1236 Researchers should ensure that curation and transformation processes do not alter the meaning of
1237 data or cause the loss of important contextual information. Descriptions of processes should
1238 include safeguards or checks to ensure that patient data are not duplicated or overrepresented. In
1239 addition, documentation of processes used to mine and evaluate unstructured data should
1240 describe the techniques employed (e.g., natural language processing) to abstract unstructured
1241 data (e.g., clinician notes) and supplement structured data (e.g., diagnostic codes).
1242

1243 Processes used for managing and preparing the final study-specific analytic dataset should be
1244 described in the study protocol or analysis plan. Analysts should have appropriate training or

Contains Nonbinding Recommendations

Draft — Not for Implementation

1245 experience with the data and software used to compile the analytic datasets. To facilitate FDA
1246 review, all submitted programs (e.g., those written by analysts) should be thoroughly annotated
1247 with comments that describe the intent or purpose of each data management and analysis step
1248 written in the program (e.g., annotate each data step in a statistical analysis program).
1249

1250

1251 **VII. GLOSSARY**

1252

1253 **Accuracy:** Closeness of agreement between the measured value and the true value of what is
1254 intended to be measured.¹⁷
1255

1256 **Artificial Intelligence (AI):** The science and engineering of making intelligent machines,
1257 especially intelligent computer programs (McCarthy 2007).
1258

1259 **Common Data Model (CDM):** Standardizes a variety of electronic health care data sources into
1260 a common format to ensure interoperability across all sites providing data.¹⁸
1261

1262 **Completeness:** The “presence of the necessary data” (National Institutes of Health
1263 Collaboratory 2014).
1264

1265 **Computable Phenotype:** A clinical condition or characteristic that can be ascertained using a
1266 computerized query to an EHR system or clinical data repository (including disease registries,
1267 claims data) using a defined set of data elements and logical expressions. Computable
1268 phenotype definitions provide the specifications for identifying populations of patients with
1269 conditions of interest.¹⁹
1270

1271 **Conceptual Definition:** Explains a study construct (e.g., exposure, outcomes, covariates) or
1272 feature in general or qualitative terms.
1273

1274 **Concomitant Medication:** Prescription or nonprescription drugs or supplements used
1275 concurrently with the product of interest or comparator agent.
1276

1277 **Conformance:** “[D]ata congruence with standardized types, sizes, and formats” (Daniel et al.
1278 2018).
1279

1280 **Confounder (Confounding Factor):** A variable that can be used to decrease confounding bias
1281 when properly adjusted for in an analysis. Confounding is the distortion of a measure of the
1282 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/>.

Contains Nonbinding Recommendations

Draft — Not for Implementation

1283 that influence the occurrence of the outcome. Confounding occurs when all or part of the
1284 apparent association between the exposure and the outcome is in fact accounted for by other
1285 variables that affect the outcome and are not themselves affected by exposure (Porta 2014).
1286

1287 **Continuity of Coverage:** The period of time over which a patient is enrolled in a health care
1288 system and during which any medical service or drug prescription would be captured in that
1289 health care system's electronic record system.²⁰
1290

1291 **Covariate:** A variable that is neither an exposure nor outcome of interest, but is measured to
1292 describe a population or because it may be a confounder or effect modifier to account for in
1293 study design or analysis.
1294

1295 **Cumulative Dose:** The total amount of the drug of interest (exposure) given to a patient over a
1296 specified period of time.²¹
1297

1298 **Data Accrual:** The process by which the data was collected.
1299

1300 **Data Curation:** Application of standards (e.g., Health Level 7, ICD-10-CM) to source data; for
1301 example, the application of codes to adverse events, disease staging, the progression of disease,
1302 and other medical and clinical concepts in an EHR.
1303

1304 **Data Element:** A piece of data corresponding to one patient within a data field (from Daniel, et
1305 al. 2018).
1306

1307 **Data Integrity:** The completeness, consistency, and accuracy of data.²²
1308

1309 **Data Repository:** A database that consolidates data from disparate clinical sources, such as
1310 those within an EHR system, to provide a broader picture of the care a patient has received.²³
1311

1312 **Data Transformation:** Includes data extraction, cleansing, and integration (e.g., into a CDM).
1313

1314 **Data Warehouse:** Consists of data from the data repository that has undergone data
1315 transformation and de-identification.
1316

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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

1317 **De-Identification:** The process by which personal identifiers are removed from an individual's
1318 health information.²⁴

1319
1320 **Distributed Data Network:** A network of multiple dispersed health care data sites providing the
1321 ability to query or analyze data from any or all sites.

1322
1323 **Effect Modifier:** A factor that biologically, clinically, socially, or otherwise alters the effects of
1324 another factor under study (Porta 2014).

1325
1326 **Electronic Health Care Data:** Analytic data that is an organized collection of automated health
1327 data available from computers or other electronic technological platforms.²⁵

1328
1329 **Electronic Health Record (EHR):** An individual patient record contained within an EHR
1330 system. A typical individual EHR may include a patient's medical history, diagnoses, treatment
1331 plans, immunization dates, allergies, radiology images, pharmacy records, and laboratory and
1332 test results.²⁶

1333
1334 **Medical Claims Data:** The compilation of information from medical claims that health care
1335 providers submit to insurers to receive payment for treatments and other interventions. Medical
1336 claims data use standardized medical codes, such as the World Health Organization's
1337 International Classification of Diseases Coding (ICD-CM) diagnosis codes, to identify diagnoses
1338 and treatments.²⁷

1339
1340 **Misclassification:** The erroneous classification of an individual, value, or attribute into a
1341 category other than that to which it should be assigned (Porta 2014).

1342
1343 **Missing Data:** Data that would have been used in the study analysis but were not observed,
1344 collected, or accessible. This refers to information that is intended to be collected but is absent
1345 and information that is not intended to be collected and is therefore absent.

1346
1347 **Negative Predictive Value (NPV):** The probability that a subject does not have a disease when
1348 the classification result is negative.

1349
1350 **Operational Definition:** The data-specific operation or procedure a researcher followed to
1351 measure constructs in a particular study.

1352

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/coveridentities/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)

Contains Nonbinding Recommendations

Draft — Not for Implementation

1353 **Plausibility:** The believability or truthfulness of data values (Kahn et al. 2016).

1354

1355 **Positive Predictive Value (PPV):** The probability that a subject has a disease when the
1356 classification result is positive.

1357

1358 **Provenance:** An audit trail that “accounts for the origin of a piece of data (in a database,
1359 document or repository) together with an explanation of how and why it got to the present
1360 place.”²⁸

1361

1362 **Sensitivity:** The probability that a classification result will be positive when the subject has the
1363 disease.

1364

1365 **Source Data:** All information in original records and certified copies of original records of
1366 clinical findings, observations, or other activities in a clinical study necessary for the
1367 reconstruction and evaluation of the study. Source data are contained in source documents
1368 (original records or certified copies).²⁹

1369

1370 **Specificity:** The probability that a classification result will be negative when the subject does not
1371 have the disease.

1372

1373 **Study Period:** The calendar time range of data used for the study (Wang et al. 2017).

1374

1375 **Traceability:** Permits an understanding of the relationships between the analysis results (tables,
1376 listings, and figures in the study report), analysis datasets, tabulation datasets, and source data.³⁰

1377

1378 **Validation:** The process of establishing that a method is sound or that data are correctly
1379 measured, usually according to a reference standard.³¹

1380

1381

1382 **VIII. REFERENCES**

1383

1384 Bustamante, R, A Earles, JD Murphy, AK Bryant, OV Patterson, AJ Gawron, T Kaltenbach,
1385 MA Whooley, DA Fisher, SD Saini, S Gupta, and L Liu, 2019, Ascertainment of Aspirin
1386 Exposure Using Structured and Unstructured Large-scale Electronic Health Record Data, *Med*
1387 *Care*, 57:e60–e64.

1388

1389 Carreras, G, M Simonetti, C Cricelli, and F Lapi, 2018, Deterministic and Probablistic Record
1390 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.

Contains Nonbinding Recommendations

Draft — Not for Implementation

1391
1392 Curtis, M, SD Griffith, M Tucker, MD Taylor, WB Capra, G Carrigan, B Holzman, AZ Torres, P
1393 You, B Arnieri, and AP Abernethy, 2018, Development and Validation of a High-Quality
1394 Composite Real-World Mortality Endpoint, Health Services Research, 53(6)Part I:4460-4476.
1395
1396 Daniel, G, C Silcox, J Bryan, M McClellan, M Romine, and K Frank, 2018, Characterizing
1397 RWD Quality and Relevancy for Regulatory Purposes, Duke Margolis Center for Health Policy,
1398 accessed January 9, 2019,
1399 https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf.
1400
1401 Girman, CJ, ME Ritchey, W Zhou, and NA Dreyer, 2019, Considerations in Characterizing
1402 Real-World Data Relevance and Quality for Regulatory Purposes: A Commentary,
1403 Pharmacoepidemiol Drug Saf, 28(4):439–442.
1404
1405 Harron, KL, JC Doidge, HE Knight, RE Gilbert, H Goldstein, DA Cromwell, and JH van der
1406 Meulen, 2017, A Guide to Evaluating Linkage Quality for the Analysis of Linked Data, Int J
1407 Epidemiol, 46(5):1699–1710.
1408
1409 Haynes, K, 2019, Mortality: The Final Outcome, Pharmacoepidemiol Drug Saf, epub ahead of
1410 print Jan 31, 2019, doi: 10.1002/pds.4715.
1411
1412 Kahn, MG, TJ Callahan, J Barnard, AE Bauck, J Brown, BN Davidson, H Estiri, C Goerg, E
1413 Holve, SG Johnson, ST Liaw, M Hamilton-Lopez, D Meeker, TC Ong, P Ryan, N Shang, NG
1414 Weiskopf, C Weng, MN Zozus, and L Schilling, 2016, A Harmonized Data Quality Assessment
1415 Terminology and Framework for the Secondary Use of Electronic Health Record Data,
1416 EGEMS, 4(1):1244.
1417
1418 Lash, TL, MP Fox, and AK Fink, 2009, Applying Quantitative Bias Analysis to Epidemiologic
1419 Data, New York (NY): Springer.
1420
1421 Mahendraratnam, N, C Silcox, K Mercon, A Kroetsch, M Romine, N Harrison, A Aten, R
1422 Sherman, G Daniel and M McClellan, 2019, Determining Real-World Data’s Fitness for Use and
1423 the Role of Reliability, Duke Margolis Center for Health Policy, accessed July 24, 2020
1424 https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf.
1425
1426 McCarthy, J, 2007, What Is Artificial Intelligence?, John McCarthy’s Home Page, updated
1427 November 12, 2007, <http://www-formal.stanford.edu/jmc/index.html>.
1428
1429 Miksad, RA, and AP Abernethy, 2018, Harnessing the Power of Real-World Evidence (RWE):
1430 A Checklist to Ensure Regulatory-Grade Data Quality, Clin Pharmacol Ther, 103(2):202–205.
1431
1432 National Institutes of Health Collaboratory, 2014, Assessing Data Quality for Healthcare
1433 Systems Data Used in Clinical Research, accessed August 27, 2019,
1434 [https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-
1435 quality_V1%200.pdf#search=Assessing%20data%20quality](https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf#search=Assessing%20data%20quality).
1436

Contains Nonbinding Recommendations

Draft — Not for Implementation

- 1437 Navar, AM, ED Peterson, DL Steen, DM Wojdyla, RJ Sanchez, I Khan, X Song, ME Gold, and
1438 MJ Pencina, 2019, Evaluation of Mortality Data from the Social Security Administration Death
1439 Master File for Clinical Research, JAMA Cardiol, epub ahead of print Mar 6, 2019, doi:
1440 10.1001/jamacardio.2019.0198.
1441
1442 Porta, M, 2014, A Dictionary of Epidemiology, Sixth Edition, New York (NY): Oxford
1443 University Press.
1444
1445 Richesson, RL, MM Smerek, and CC Blake, 2016, A Framework to Support the Sharing and
1446 Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research
1447 Applications, EGEMS, 4(3):1232.
1448
1449 Szklo, M, and FJ Nieto, 2006, Epidemiology: Beyond the Basics, 2nd Edition, Burlington (MA):
1450 Jones & Bartlett Learning.
1451
1452 Velentgas, P, NA Dreyer, P Nourjah, SR Smith, and MM Torchia, editors, 2013, Developing a
1453 Protocol for Observational Comparative Effectiveness Research: A User's Guide, AHRQ
1454 Publication No. 12(13)–EHC099, accessed January 9, 2019,
1455 [https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/user-guide-observational-cer-
1456 130113.pdf](https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/user-guide-observational-cer-130113.pdf).
1457
1458 Wang, SV, S Schneeweiss, ML Berger, J Brown, F de Vries, I Douglas, JJ Gagne, R Gini, O
1459 Klungel, CD Mullins, MD Nguyen, JA Rassen, L Smeeth, and M Sturkenboom, 2017, Reporting
1460 to Improve Reproducibility and Facilitate Validity Assessment in Healthcare Database Studies
1461 V1.0, Pharmacoepidemiol Drug Saf, 26(9):1018–1032.