ISPOR
The professional society for health economics and outcomes research

Improving healthcare decisions

505 LAWRENCE SQUARE BLVD SOUTH
LAWRENCEVILLE, NJ 08648

P +1-609-586-4981
F +1-609-586-4982

info@ispor.org
www.ispor.org

November 29, 2021

Docket Number: FDA-2020-D-2307

Dear FDA:

ISPOR – the professional society for health economics and outcomes research - is pleased to respond on behalf of its membership to your consultation entitled "**"Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products."**

ISPOR is a scientific and educational society with many of its members engaged in evaluation of health technologies, including pharmaceuticals, medical devices, and other interventions. We have a large membership living and working in 110 countries globally, across a range of disciplines, including health economics, epidemiology, public health, pharmaceutical administration, psychology, statistics, medicine, and more, from a variety of stakeholder perspectives, such as the life sciences industry, academia, research organizations, payers, patient groups, government, and health technology assessment bodies. The research and educational offerings presented at our conferences and in our journals are relevant to many of the issues and questions raised in this request for information.
The response to this consultation was led by the Policy Outlook Committee of our most senior advisory body, the Health Science Policy Council. To engage our membership, we consulted with interested members of our Real World Evidence Steering Committee, Institutional Council (ie, industry and consulting), and Faculty Advisor Council, and solicited our general membership for comments. The attached document provides a synthesis of their comments. We hope they prove useful.

ISPOR would be happy to answer any questions about our response, as well as to participate in any follow-up consultations on the relevant program items mentioned within the report.

Sincerely,

Nancy S Berg

Nancy S. Berg
CEO & Executive Director
ISPOR

ISPOR Comments on FDA Draft Guidance "Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products", September 2021. (Docket Number: FDA-2020-D-2307)

General Comments on the Draft Guidance

Overview

The draft guidance document is a very useful resource about the level of transparency FDA seeks from sponsors to evaluate electronic health records and claims data. We look forward to the additional guidance regarding study design and analysis and regulatory considerations in fulfillment of PDUFA and Cures commitments for FY2021.

Given the connections between the four proposed draft guidance documents related to RWD/RWE planned for release, it would be useful if FDA could keep the comments period open for all the draft guidance until the final submission period for the fourth draft guidance.

We would also ask the FDA to recognize that implementation of these standards will necessarily change over time as the FDA gains experience with specific data sources and as data sources evolve.

Need for collaboration

There are gaps in this guidance that can be addressed with more detail and evaluation of growing evidence to support the value of using different RWD sources, challenges of study design elements, and data quality as it relates to the use of RWD within the medical and clinical research communities. The gaps identified in this guidance validate the suggestion that the FDA form and organize working groups with industry, digital data technology, data vendor companies, life sciences researchers, bioinformatics, ethics, and data privacy stakeholders and experts to create focused guidance and provide tools that create a learning healthcare system necessary to transform healthcare using RWD. There is still much to be learned and the FDA's role in creating informed, supportive, and impactful practices for the use of RWD for regulatory decision making requires continued collaboration and input from stakeholders to support the FDA as it defines guidance, and policies, that provide the FDA with the right data, for the right purpose, the right methods, the right time, and the right patient/populations, for who regulatory decisions will be made using RWD.

There are some areas of collaboration across stakeholders that are immediately apparent. To support transparency in reporting about data aggregation, curation, and provenance will require cooperation of data aggregators/curators and the researchers that create the analytic data set and execute the study protocol. Specifically, the information requested in sections IV, V, and VI is very extensive, and it is the data providers—not the sponsors—who can provide many of these details. It may be more feasible and efficient for data providers to prepare a summary document with these details that could be included in the appendix of study protocol/report. In addition, given the time and resource necessary to conduct validation studies for exposures, outcomes, and covariates in a given study, it would be useful to establish ways to enable and encourage sharing of validation study results across study sponsors to allow for more efficient study execution.

Prespecification and preregistration

Prespecification of protocols and analysis plans is critical and we are happy to see that called out in the guidance. The Real-World Evidence Transparency Initiative's registry, a partnership between ISPOR, ISPE, Duke-Margolis, and the National Pharmaceutical Council, may provide a useful vehicle for this information in RWE trials. This registry has an option to keep the pre-registered protocol in a lockbox, available only to specified reviewers and not the general public for a given period of time. Specified reviewers may include regulatory personnel.

Process

We would encourage FDA to address in more detail the process for sponsors to engage the agency in discussions for using RWD/E to inform a regulatory decision. We appreciate the note in the General Considerations section that sponsors should submit protocols and analysis plans and request comments or a meeting if seeking FDA input prior to conducting the study. FDA expectations for sponsor interactions with the review divisions could benefit from more details with respect to process, type(s) of meeting, suggested timing/timelines, and anticipated role of the RWE Subcommittee. In addition, we suggest the FDA consider a more formal independent review committee who can bring appropriate expertise to all submissions involving RWD/E across review divisions and thereby ensure greater consistency in process.

During the protocol development stage, the sponsor often does not have access to different data sources to address all the key considerations described in the draft guidance. Is there any distinction between information expected for evaluation during the protocol development stage versus during the study conduct and final report?

Fit for purpose

The guidance reads as if all RWD needs to meet the same bar, regardless of whether the information is person-generated, health insurance claims aggregated across many insurers, or individual EHR records downloaded with patients' permission. For example, we have seen important regulatory approvals that were based in part by benchmark or comparison RWD. In some of these situations, these data may not pass the highest quality bars, but are nonetheless indicative of the natural history of the disease and the notable difference in outcomes between a new medication and those untreated has been sufficiently persuasive to warrant a new approval or expansion of indication. Similarly, the pandemic has shown the regulatory value of using new data sources, such as person-generated health data. This could be a good example of recognizing the value of such information, which could be further enhanced by clinical validation which could range from reviewing the person's electronic health record or checking medical claims to confirm that indeed a doctor visit occurred on the date reported by the patient for an event consistent with that reported by the patient. It is very encouraging that FDA does not seek to limit the possible data sources that may be relevant to answering study questions.

Hence, data relevance and reliability must be considered on a spectrum. FDA should recognize that there will be a continuum of data quality across data sources and that trade-offs may be needed, especially where there may be limited data sources to study rare conditions or emerging endpoints within specific indications. The point here is that data relevance and established quality may not have complete congruence, and highly relevant data that has not undergone formal validation may nonetheless be very informative by providing essential benchmarks

Accordingly, the guidance should incorporate the key validity concepts and FDA definitions of "fit for purpose" and "context of use." Similar to how these terms are being incorporated in the new guidance on validity of Clinical Outcomes Assessment instruments, they should also be incorporated in guidance on validity of operational definitions derived from EHR and claims data. In particular, "fit for purpose" is defined as "a conclusion that the level of validation…is sufficient to support its proposed use." "Context of use" is defined as "a statement that fully and clearly describes the way the medical product development tool [or operational definition] is to be used and the regulated product development and review-related purpose of the use.

It would be helpful for FDA to provide a framework on different regulatory contexts of use for observational studies and how 'fit for purpose' validation of operational definitions would be considered within each context. Sponsors would benefit from additional guidance when evaluating EHR and claims data sources to determine when the data are adequate for the intended use.

Summary/checklist

The Draft Guidance covers a broad range of topics and recommendations in great detail. It would be helpful if the FDA would classify and summarize both the minimum requirements at a high level as well as potential enhancements that would strengthen the evidence quality. A checklist of key recommendations would be helpful.

Provision of additional information

We recommend that FDA reference key scientific papers and authoritative perspectives (e.g., ISPOR, ISPE, ENCePP, others) that discuss use of EHRs/claims and how to overcome limitations to complement to the examples provided.

Also, while the document is a comprehensive, useful resource, it might be helpful to remove some extraneous background and basic educational information (e.g., details on distributed data networks) to streamline the document.

Below, we provide comments on the three major sections, IV, V, and VI.

Section IV. Data Sources

General section comments

This section is comprehensive. First we offer a few section-level points where clarification would be helpful, and then proceed to line-by-line comments for each section.

In Section 1, on enrollment and comprehensive capture of care, EHR and claims data are often described with the same broad brush. There are several areas where important differentiations could be called out, such as enrollment and disenrollment which are not applicable to EHRs in most instances. Another example is stating the capture of nonprescription or drugs that are not reimbursed under insurance plans are not systematically captured in either data source. While factually true, EHRs often contain nonprescription medication use reported by the patient and written prescriptions for drugs regardless of prescription plan coverage. Finally, the FDA might consider including a distinction between data source completeness and comprehensiveness.

In Section 2 - data linkage and synthesis, many data source owners or aggregators are offering deterministic linkages from EHR/claims to specialized registries, mortality records, and social determinants of health. These linkages are often predetermined and 'embedded' within the data offering. RWD owners and aggregators will be responsible for enumerating the information requested here. There may be an opportunity to prequalify certain linked datasets or linkage methods such that these questions can be answered by specific DSOs/aggregators. For distributed data networks, while common data models can allow for disparate data to be aggregated based on certain data elements, each data source is unique and therefore applying a CDM to every dataset may create situations where some nuanced data available only in some sources could be lost to the detriment of the study hypotheses. Therefore, we would urge caution in applying a CDM in every situation or at least allow for data capture outside a CDM where key data elements may reside.

In Section 5 - unstructured data, there is no guidance for the situation where the process of NLP and other AI to be proprietary to the developer/owner and considered intellectual property.

B1. Enrollment and Comprehensive Capture of Care (Lines 201-247)

Overall, this section describes the many ways claims or EHR data may not be fit-for-purpose. In addition, FDA should provide guidance on what would be considered acceptable in a possible RWD source qualifying it as fit-for-purpose rather than focusing only on what disqualifies it. Comprehensive capture of care does not always equate to care in all settings. Specialty EMR data, for example from oncologists, may be more comprehensive in describing and evaluating certain exposures and outcomes than an EMR that is more general in design. One example is that an oncology-specific EMR could explicitly capture lines-of-treatment where development of rules to define a line of therapy may have to be defined by the sponsor and implemented using a programmatic algorithm.

Enrollment (lines 201-210) is not generally applicable to EHRs. Proxies usually involve assuming that a patient is eligible to receive care between any two dates of service in the EHR. The patient is then censored at the last date of service recorded. In addition, patients tend to stay longer with their healthcare providers than any insurance provider either through a change of employment or change of in life circumstances (e.g, aging into Medicare). Specific to claims data, perhaps the guidelines should explicitly state that a claims database is required to have enrollment records for use in evidence generation activities for the FDA. If, however open claims (provider/clearinghouse sourced claims data) are to be considered by FDA for regulatory decision-making then this requirement is not so black and white.

In the next paragraph (lines 212-219), the FDA puts the onus on the sponsor to report "the likelihood that all exposures and outcomes of interest will be captured for regulatory decision-making". The sponsor may not know "all" that FDA will consider in their decision process. The current guidance includes an example where an outcome is dependent on a "specific frequency of laboratory tests, and clinicians do not typically order those tests at such a frequency". The point we believe the FDA guidance is making is that care in standard medical practice may not match that outlined in a clinical trial. One example is the grading of adverse events which providers in the "real-world" do not assess. In RWD this may be proxied based on the setting or type of care provided such as an emergency department visit or hospitalization to be equivalent to grade 3 and 4 events.

Lines 223-224, the FDA might consider defining how sponsors could evaluate a data source contain "adequate" sample size and length of follow-up. Assessment of something being "adequate" is subject to interpretation and the sponsor and FDA may differ in what is deemed as such.

Lines 231-233. FDA should consider updating the sentence to note that EHR may capture non-prescription drugs (patient reported medications) and drugs not reimbursed under health plans (prescriptions written) while claims data will not.

B2. Data Linkage and Synthesis (Lines 248-289)

Some data may be obfuscated following expert attestation of de-identification. For example, date of death may be truncated to month and year of death, and geographic location may be limited to a census region. FDA should note that here as well, the protocol should describe how this will be handled for specific aspects of the analysis

Linkage, by definition, creates the intersection of linked datasets if only linked records are retained. This can have a strong impact on the patient population represented in the data. For example, claims data are payer specific (e.g., Medicare versus a commercial health plan). An EHR may initially reflect all age groups but, upon linkage with a Medicare dataset, will mainly reflect an older population and only those enrolled in fee-for-service versus Medicare-Advantage plans.

When linking data sets together, information may exist in both which may or may not agree in value. For example, as EHR is meant to track patients' health and healthcare, claims are meant for billing purposes. ICD-10 diagnosis codes have different levels; what is recorded in the EMR may be more or less refined in the claims for the purposes of reimbursement. Guidance on how sponsor should provide justification for the decisions and possible ways of assessing the impact of these decisions (e.g., assessing degree of agreement or conducting sensitivity analyses) would be useful.

FDA should provide guidance for when sponsors obtain linked data from data aggregators but these service providers do not provide all the detail in how the linkage is performed. At a minimum sponsors should provide the name of the data aggregator, any rules (e.g., tokens) used for matching and the reported accuracy of the chosen combination from the aggregator.

When data sources are linked, the resulting database must go through a formal de-identification attestation by a qualified expert as defined in Section §164.514(b)(1) of the HIPAA Privacy Rule. While individually databases may be attested as de-identified, once combined the additional information contained in the other source may negate that determination. Hence individual measures or patients may be subject to additional obfuscation. For example, extremes in body weight (low and high) might be subject to truncation, or age might be truncated 80 years of age where in the individual databases it was 90 years of age. Other measures may need to be removed in their entirety or dates, such as date of death may need to be reported as month and year, with the actual day removed. The guidance should recommend the sponsor evaluate whether the linked data remains fit-for-purpose or whether the use of each source individually may be more appropriate and document any known obfuscations which may impact the measurement of the exposures and outcomes.

Lastly the FDA recommends documenting what curation may be performed to address duplication when linking or combining data sources in other ways (lines 285-288). The FDA should take note that some data owners will restrict a licensee from trying to remove duplicate patients or records. Identification of such, depending on the method, is seen as a re-identification process and violates the de-identification attestation of the combined data.

B3. Distributed Data Networks (Lines 290-355)

Federated networks should also be included in this section, as mapped and combined data are not stored in a centralized repository until such time as patient selection and analysis is conducted. Often the data is extracted and mapped in real-time. The primary benefit of a federated network is to increase sample size and power, not to execute identical queries on multiple datasets as the guidelines state is the case with distributed data networks.

B5. Unstructured Data (lines 369-394)

With respect to describing the process of NLP or other AI implementation, the reporting of this data into structured fields may be provided by a data vendor such that the licensee may not have access to the scripts used or the validation results (lines 383-393). There is no guidance for the situation where the process of NLP and other AI to be proprietary to the developer/owner and considered intellectual property. What guidance does the FDA have in these circumstances.

Guidance on describing data abstraction processes from unstructured data in patient medical charts that is not technology assisted should be provided. Whether it is agreement of data capture through abstraction of the same records by two different abstractors (e.g., kappa statistics and acceptable levels of agreement), checks put in the electronic data capture tool and quality checks following abstraction (e.g., death date before birth date, incorrect unit of measurement such as temperature recorded as Celsius when it was in fact Fahrenheit. The FDA might consider requesting the abstractor training materials and anything associated with the criteria to apply during eCRF build.

C. Missing Data: General Considerations (lines 395-423)

Data linkage may identify significant additional missing data apart from item missingness. In particular, linkage of claims and EHR data will generally identify substantial portions of the patients with enrollment and healthcare utilization in the claims data with no accompanying EHR records or the reverse. It may also identify services missing from the EHR record but included in the claims data for overlapping periods. This is also a good check for how complete an EHR might be for evaluation and generation of evidence for a specific disease or population under study.

D. Validation: General Considerations (lines 424-499)

The draft guidance places significant emphasis on the role of formal validation across outcome, exposure, and covariates with focus on validation against source medical records or cross-validation using linked datasets (EHR and claims). While this best practice is laudable, FDA should recognize that this (1) is not always feasible depending on the data access and governance model, and (2) may have the unintended consequence of limiting the scope of RWE generation to data sources with certain access, data governance, and staff resource models.

Further clarity on the FDA's expectations for validation and use of novel or small datasets is requested as these may be the only sources of RWD available when seeking approval under the Orphan Drug Act of 1983 for rare diseases. Would the FDA to consider more practical approaches to validation, such as example situations in which use of previously published validation work using the same or comparable data source would be acceptable? We do propose that a sponsor start with more general data characterization before prioritizing more in-depth validation of key variables (as possible for a given data source) with elements from the analysis plans.

l. 444. Complete verification: Suggest that FDA acknowledge that chart review studies based on electronic health records and EHR databases that use chart review (rather than relying on structured fields) would meet the definition of complete verification.

V. Study Design Elements

General section comments

This section provides a detailed review of design considerations pertaining to observational studies using RWE. However, in some areas, the guidance is so high-level and the examples are so specific that the guidance does not cover the breadth and depth of design considerations necessary to truly serve as a roadmap for product manufacturers and RWE scientists. Our detailed comments on this section are summarized below and may be useful in refining the draft guidance.

We noted that the draft is particularly helpful in pointing out many of the pitfalls in designing RWE research, which are helpful to be aware of up front when designing the study. In an ideal scenario, all of these pitfalls would be addressed-- but given that all research designs involve tradeoffs, it is likely that some but not all can be addressed. This point is important since raising the bar too high may leave few studies suitable for meeting the criteria used in regulatory decision making.

One notable opportunity for improvement in this section is clarification as to how randomized controlled trials (RCTs) conducted for the product of interest may serve to inform RWE design. For example, the time horizon employed to observe the study outcome in the trial(s) may also be appropriate to the time horizon for measuring the outcome in a RWE study. Also, the treatment regimens and patient populations used in the RCTs may be helpful to informing these elements of RWE design. Or, conversely, the RCT design or results may have gaps that can be addressed by RWE observational research—in which case it is important to identify the gaps and how each will addressed by the proposed RWE study.

Also, given that this guidance is geared towards comparative effectiveness and safety studies, RWE study design could follow a target trial approach, proposed by Hernan and Robins 2016. The general idea is to apply design principles from RCTs to observational studies by specifying the hypothetical RCT (the 'target trial') that would answer the question of interest (e.g. treatment effect or adverse event).

Finally, one of the crucial study design elements that is missing from the draft guidance (and often overlooked in practice) is the alignment between patient eligibility, baseline and treatment initiation (beginning of follow-up). Failure to properly define this will introduce bias, particularly in RWE studies where eligibility criteria can be met at multiple time points.

We offer the following specific comments about this section:

Lines.503-504. We agree that the research questions shall be defined in advance and addressed independently from the data sources. However, this does not need to happen though a single study. The design of a single study can therefore be affected by the data sources, and indeed, it is often the case. For example, consider a research question being the evaluation of the extent of use of a product by pregnant women and the extent of congenital malformations in the offspring of those who are exposed to the drug. Initially the researchers may consider conducting a study addressing the entire research question. After a preliminary assessment of data sources, the researchers may decide to split the study into two: a drug utilization study using a wide range of data sources capturing the exposure among pregnant women, and a cohort study ascertaining the outcome in the offspring using those data sources which provide the mother-baby linkage as well as the outcome in the offspring.

Lines 506-508. We suggest that should also be discussed among the study design elements. The choices and techniques of matching or even pseudo-randomization shall be part of this guidance as it raises considerable discussion at the time of the designing of studies.

Line 547. Minor comment – the document makes reference to International Classification of Diseases, Ninth Revision, Clinical 547 Modification (ICD-9-CM) diagnosis codes. This could be updated to ICD-10 to reflect more recent coding practices.

Line 560. Current guidance focuses on the definition and validation of 'exposure' but provides little insights on the choice of comparators (section 7 is insufficient). Comparator data might not even come from the same study. For example, uncontrolled studies, such as single arm trials, are often used to inform regulatory decisions and might require the use of external controls.

Lines 668-670. Sentence says: *"Other than for medications administered in hospital settings or infusion settings, electronic health data capture prescriptions of drugs and the dispensing of drugs to patients, but generally do not capture actual patient drug exposure because this depends on patients obtaining and using the prescribed therapy."* It should also mention medications administered in outpatient settings other than hospital or infusion centers (examples of products administered in such outpatient settings include corticosteroid injections and vaccines). Sentence could be revised to "Other than for medications administered in hospital settings, infusion settings or outpatient practices…"

Lines 720-725. Relevant concomitant medications should be elaborated on a bit here—most relevant are those that can affect either the likelihood of the study outcome or exposure to the drug being studied (the latter would be the case if the concomitant drug interacts with the study drug such that taking the two together could impact the absorption, distribution, metabolism, and/or excretion of the drug being studied).

Line 727. EHR databases that use technology-enabled abstraction have already conducted medical record review for key variables using standardized processes, i.e. complete verification has been performed based on the information available in the patient chart. Does FDA agree that variables abstracted in this way have been 'validated' via complete verification, provided that sponsors provide descriptions of the medical record review processes, kappa statistics, etc.?

Line 727. For some endpoints, such as laboratory-based measures (obtained from structured, normalized lab data, e.g., HbA1c, neutrophil count, creatinine) or events derived from unstructured data (e.g. tumor response or progression), there is no clear reference standard available. Additional guidance is needed from FDA regarding acceptable approaches to validate these types of real-world endpoints as well as what supportive data could be provided to increase confidence in these variables. The guidance is focused on criterion validity and analytical validation; however, these endpoints may need to be assessed in other ways that demonstrate clinical validity, such as considering face validity with experts, evaluating the completeness of the underlying data, measuring reliability between abstractors, and evaluating performance in terms of correlation with other related outcomes.

Line 727. This section focused on discrete outcomes or acute events as outcomes and mentioned Mortality as an outcome. For completeness, this section may need to add continuous variables as outcomes (e.g. laboratory values) as well as other time-to-event variables (e.g. progression-free survival) as outcomes. For time-to-event outcomes, censoring should be addressed.

Line 784. We suggest that this section of the guidance reference current Clinical Outcome Assessment (COA) guidance on validation of patient- or physician-generated data to assess outcomes.

Line 836. Regarding the medical record retrieval rate discussed in this section, it's not clear what this rate would be used for.

Line 865. How is the necessary level of certainty determined for different contexts of use? An example would help sponsors understand FDA's thinking.

Line 981. This section focused on individual variable-level ascertainment and validation. There is no guidance here on the collective covariate level ascertainment and validation. The collective covariate level ascertainment and validation refers to determination of the minimal set of covariates which describe the population sufficiently and reliably; therefore, the population could be used for the purpose of interest.

Line 1001. The guidance on potential linkages with other data sources to help address unmeasured or imperfectly measured confounders is helpful. As there are many diverse research scenarios with highly varying availability of information, we recommend broadening this advice to include other sensitivity analysis methods (e.g., Zhang & Mather, 2020) that help to quantify potential impact of confounding and reduce the uncertainty to levels allowing for confident decision-making.

Line 1000-1003. Here, it states that potential linkages with other data sources or additional data collection be performed to capture important confounders but does not state the obvious which that unmeasured or imperfectly measured confounders be disclosed by the researcher.

Line 1005. Similar to the previous comment about confounders, this section on Effect Modifiers does not state that the researcher should disclose such modifiers regardless of whether they can be measured or are imperfectly measured (current guidance only advises that they are examined).

Line 1012. Biomarkers and genetic mutations are important effect modifiers that could be mentioned in the guidance.

Section VI. Data Quality During Data Accrual, Curation, and Transformation into the Final Study-Specific Dataset

General considerations

FDA provides examples of validation based on linked EHR and claims, but in many situations these data will not be available. FDA should provide additional guidance on more flexible approaches and considerations for reconciliation outside of formal linkages and key points for consideration in those circumstances.

For studies that would use data that also include pandemic time periods, due to rapid disruptions in health care delivery, are there any specific considerations should sponsors undertake?

Data sciences and artificial intelligence advances rapidly, and their use may not be restricted to "(1) extract data elements from unstructured text in addition to structured fields in EHRs; (2) develop computer algorithms that identify outcomes; or (3) evaluate images or laboratory results", e.g., they can be used to evaluate data quality and identify variable patterns. These additional uses should be added.

Specific comments

Line. 1069. Consensus based data standards – please clarify exactly what these are. In the data standards guidance, they specify CDISC and SDTM – is this what is mentioned here?

Lines 1069-71. We suggest that evaluation should be performed consistent with the importance of specific data fields (i.e. critical outcomes may need to be thoroughly assessed for completeness, accuracy, and plausibility, whereas other data elements may not require verification against source documents.)

Line 1078. It is recognized that sponsors who design and execute studies may not be directly involved in the first cycle of curation by data aggregators; adequate documentation of curation procedures should be made available by the data aggregators.

Lines 1078-1084. It would be helpful if this was split into responsibilities for sponsors and responsibilities for data aggregators – a flow chart might be helpful to do this. For example, it is expected that the reports on data accrual, creation, and transformation will likely be produced by data aggregators/curators, whereas reports on study-specific analytic data set creation as well as QA/QC documentation of the analytic data set will likely be conducted by those who design and execute the study protocol.

Line 1084. Reference to “researchers” - would be good to specify the role of data aggregators here rather than just broadly defining “researchers”.

Line 1152. It is good to see opportunity/acknowledgement to the use of “foreign data” alongside US data

l. 1159-1160. Re-identification of patients might not be possible in data sources in general and regulations may prevent this in other countries – could the FDA provide clarification on this?

Related to above, the wording of Section VI appears to limit the use of data that is not directly available from the data owner. Some data providers license their data from multiple data owners and are not involved in the initial data collection steps. Related to this point, can FDA please clarify this statement in the Data transformation section (lines 1159-1160): “De-identification of patient records and *ability to re-identify unique patients* in original source data without losing traceability.” Many vendors are unable to identify patients and/or prohibited from re-identifying patients in the data due to privacy rules.

l. 1162-1163. Many of these steps are completed prior to sponsors accessing the data for research, and some of the recommended requirements would require data providers to share proprietary information. As a result, sponsors don’t have access to the level of detail recommended in this section. For example, details of data origin can be limited due to data use agreements, and details regarding algorithms for AI, ML, and NLP of unstructured data may be considered proprietary. Can FDA please comment on what sponsors should do when this issue occurs?

l. 1218-1231. FDA should more clearly define expectations for what information from sponsors should be furnished and when. For instance, the draft guidance is not forthcoming in providing specifics on documentation needed for "traceability" or "provenance". What does FDA expect to see from data providers in terms of the provenance of a data set from an event prompt to data being in a research-ready format and then to tables, figures, and listings/study report from a sponsor? Is access to data provider SOPs (e.g. for QA/QC) and data management plans plus a protocol/SAP, computer programs, and study report sufficient? Are analytic data sets expected to be handed over (this could be challenging in many instances where sponsors do not own data or have license to do so)? It would also be helpful for FDA to clarify whether this extends to analytic tools (e.g. Panalgo IHD, Aetion, Safety Works, among others) which are increasingly being used by sponsors for analysis of RWD. It is important for sponsors to have clarity on these expectations so that they can evolve their internal process accordingly (if needed).

l. 1232-1250. FDA should recognize that the level of documentation around data accrual and curation will vary across data sources and datasets. Flexibility will initially be needed to prevent inadvertent bias towards larger data providers and inadvertent restrictions on the ability to study some disease areas that may rely on data from specialist centers (outside of broader EHR networks) which do not already have in place the infrastructure to meet the expectations for data accrual and curation documentation as set forth in the draft guidance.