January 24, 2022

Dockets Management Staff (HFA-305)
Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852

Re: Docket No. FDA–2020-D-2307: Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products

Dear Sir/Madam:

The Biotechnology Innovation Organization (BIO) thanks the Food and Drug Administration (FDA or Agency) for the opportunity to submit comments regarding the Draft guidance on **Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products.**

BIO is the world's largest trade association representing biotechnology companies, academic institutions, state biotechnology centers and related organizations across the United States and in more than 30 other nations. BIO's members develop medical products and technologies to treat patients afflicted with serious diseases, to delay the onset of these diseases, or to prevent them in the first place.

Overall, the draft guidance provides sponsors with helpful clarifications about the Agency's expectations for providing high quality real-world data that are reliable and relevant for various types of clinical studies. BIO is pleased that the draft guidance underscores the principle of transparency in conducting real world data and evidence (RWD/E) studies that the entire research community can be held to, including pharmaceutical companies, health technology organizations, regulators, and academics. BIO applauds the level of detail within this guidance in setting clear parameters for the design and delivery of RWE for regulatory decision making. Harmonization on the use of RWD/E for regulatory decision making across all FDA product Centers is of paramount importance. BIO welcomes the opportunity to discuss these comments and recommendations further with the Agency.

BIO has 3 key overarching recommendations to improve this draft guidance and bring further clarity to the use of real-world data and evidence for regulatory decision making:

1. **Incorporate appropriate degree of regulatory flexibility:** The draft guidance seems to focus almost exclusively on studies aimed at determining causal inference and does not underline the need for a fit-for-purpose approach to RWD (e.g., RWE studies may also be descriptive and may provide supportive or contextual evidence for a particular regulatory decision, contributing to the totality of evidence). BIO requests that FDA incorporate an appropriate degree of regulatory flexibility into the draft guidance that is tailored to the specific context for the RWD, and how it can support a regulatory decision as part of a totality of evidence. This would allow for the use of RWD as fit-for-purpose in filling critical

evidence gaps in benefit-risk and/or safety assessments. For example, this would include an approach tailored for aggregated data and data used to establish natural history, standard of care, or to add context to the results of an RCT. Such an approach would also include an approach to validation that prioritizes key variables needed to answer the study questions rather than all covariates. BIO also recommends that the Agency consider highlighting the use of prospective observational studies within the draft guidance and consider linking to the draft guidance on the estimand framework.

When discussing the reliability and relevance of data, BIO recommends that the Agency offer considerations for determining whether the data are fit for purpose, i.e., sufficient to support the intended regulatory context of use and should reflect the current data landscape for different disease areas recognizing that in some disease areas, for example, rare or life-threatening diseases, there may be limited data options which require more flexibility in the 'fit for purpose' scale. Sponsors would benefit from additional draft guidance when evaluating EHR and claims data sources to determine when the data are adequate for the intended use.

2. **Identify streamlined and efficient FDA-sponsor communication methods to facilitate rapid evidence generation:** Sponsor engagement with the Agency has required multiple meetings per project and the type of protocol and SAP submission process envisioned under the draft guidance could take several months-to-years to develop per research project. We are hopeful that the structured RWE pilot program and new Type D meeting under PDUFA VII may help to streamline this type of scientific dialogue, while further publication of RWE draft guidance's and case-studies may address other operational issues without necessitating multiple formal meetings. BIO proposes that the Agency work with sponsors to ensure that FDA-sponsor communications are streamlined and efficient to facilitate rapid evidence generation. Specifically, it would be helpful to provide more clarity on the amount and types of information required to have a robust/productive scientific meeting with FDA.

   BIO also recommends that the Agency provide further clarity on how FDA RWE subject matter experts, including members of the RWE Subcommittee and the Office for Biostatistics and Epidemiology, will be included in sponsor discussions with therapeutic review divisions concerning use of RWD/E. This is important to ensure that relevant expertise is consistently incorporated in agency-sponsor communications and in submission reviews. BIO also recommends that the Agency consider mechanisms to include data curators, i.e., providers of RWD/E, to discuss topics on RWE platform development unrelated to a specific sponsor study or use case of RWE.

3. **Identify best practices for data curation, processing and governance:** For a variety of reasons, including data privacy requirements, it will not be feasible for sponsors to access the entire flow of patient data as they typically license data from third parties. To the extent possible, BIO recommends that FDA work with stakeholders to better understand and address the practical realities of data curation, processing, and governance. FDA is in a unique position to bring all of the relevant stakeholders together to understand the expectations for RWD quality, and to surface and start to address challenges so that sponsors can help meet the expectations. BIO encourages FDA to work directly with health technology organizations that generate RWD and host workshops with all stakeholders to better understand these issues and help identify ways to meet Agency expectations.

Additional Considerations

1. **Scope of Draft guidance**

   - The draft guidance has explicit focus on EHRs and claims data only. It seems, however, that a lot of the content could apply to RWD in a broader context (i.e., registries, chart review studies, prospective observational studies, and studies that involve primary data collection). There are other types of "electronic health care data", e.g., data from personal electronic wearables, patient reported outcomes (PROs) in electronic diaries, clinic-genomic data etc. The draft guidance appears to only cover claims data generated by inpatient and outpatient services covered under medical benefits, but pharmacy/prescription claims are an essential part of insurance claims data. In addition, there are other forms of data such digital health technology and social media. BIO recommends that the Agency expand the guidance to include these types of data.

2. **Expectations for Protocol and Statistical Analysis Plan**:
   - The draft guidance is explicit in the types of information, justifications, and descriptions that are expected to be included in the protocol. While many of these topics would be expected for a typical clinical research protocol, we note that the expectations in the draft guidance are extensive and depending on how it is implemented may make the RWE study infeasible. To the extent possible, BIO encourages a streamlined approach to protocols for non-interventional RWE research studies rather than applying a framework better suited for clinical trials. BIO also recommends that the Agency emphasize a fit-for-purpose approach to requirements that is commensurate with the intended use of the RWD and regulatory decision to be made. Lastly, BIO recommends that the Agency clarify:
     - which information FDA expects to be in materials that should be pre-specified and submitted to the Agency prior to conducting the study (i.e., Protocol and Analysis Plan), versus
     - information that could go into materials that would only be provided to FDA after study completion (i.e., Study Report), versus
     - the information that only needs to be available upon request in case of an inspection or audit by the Agency (i.e., data quality reports, etc.).

     It would be helpful for the Agency to consider adding an appendix, or a Q&A companion document generated to specifically summarize/extract all of these recommendations for what should go into a protocol, SAP, Study Report, etc., into a single place.

3. **Nomenclature and Consistency of Definitions**

   - The draft guidance refers to one source of real-world data as "Medical Claims Data." This nomenclature is confusing as "medical claims data" in traditional Health Economics and Outcomes Research (HEOR) refers to claims submitted by physicians, hospitals,

and other providers. In reading the entire draft guidance it seems the intent is to include both pharmacy and medical claims data. BIO recommends that the Agency change the nomenclature of the guidance from "Medical Claims Data" to either "Medical & Pharmacy Claims Data" or just "Claims data" to reflect the full range of claims data used in research on biopharmaceuticals.

- BIO recommends that the draft guidance 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 draft guidance on validity of Clinical Outcomes Assessment instruments, they should also be incorporated in draft guidance on validity of operational definitions derived from EHR and claims data. For example:
    - Fit for purpose = "A conclusion that the level of validation…is sufficient to support its proposed use."
    - Context of use = "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.
- BIO also recommends that the Agency 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.

4. Additional RWE Draft Guidance:

- BIO recommends that the Agency consider providing additional guidance on:
    - obtaining consent from subjects to use their data for this kind of investigation;
    - addressing evidence generation outside pharmacoepidemiology, such as work on patient reported outcome data, indirect comparisons, and meta-analyses;
    - offering specific draft guidance on developing and validating real-world effectiveness endpoints;
    - discussing tokenization and linked datasets;
    - and providing specific considerations for certain other types of RWD.
- 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 draft 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 draft 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, and evaluating performance in terms of correlation with other related outcomes.

5. Data Quality (Including IP, Privacy and Confidentiality)

- Under General Considerations (Section III), practical recommendations to address the issue of reproducibility are needed. Many robust sources of RWE may be excluded due to regional confidentiality and patient privacy laws that limit their ability to be submitted directly to the Agency as data sets in the regulatory submission. An alternative approach for demonstrating reproducibility of the findings without directly submitting RWD as part of a regulatory submission package was recently accomplished with another regulatory body. In the absence of datasets, a clear and transparent

communication of how data were collected, curated, and analyzed is needed to ensure the quality of RWD in a regulatory application.

- In several cases, the Draft Guidance mentions that certain source data and algorithms should be provided in the protocol. This is most relevant as it pertains to AI-based extraction tools for unstructured data. These algorithms may be proprietary in nature and the expectations outlined in the guidance may necessitate data providers sharing IP with drug sponsors. BIO recommends that the Agency consider an alternative solution which would be to provide descriptions of the source data, performance characteristics of the algorithm and a description of training set data to build confidence in the extracted data. The algorithms and source data should be available for FDA's review upon inspection.
- BIO also recommends that the Agency consider developing practical recommendations to ensure transparency of methods and analysis plans through proactive stakeholder engagement within the scientific community. Accepted good practices for transparency in RWD/E study design and execution are important, such as posting the study on clinicaltrials.gov or other available forums.

6. Acceptable Data Standards and Common Data Models:

- The draft guidance further recommends the use of Common Data Models (CDMs) to harmonize the data sources. BIO recognizes the need for flexibility in use of CDMs based on the study question and the specific data structure/dataset. BIO recommends that the Agency reference existing HHS and FDA projects in this space (e.g., mapping FHIR to CDISC) and provide further details on CDMs that are already well-developed, such as OMOP/ODHSI and HL7/FHIR. BIO also recommends that the Agency clarify the level of documentation and transparency to be provided when data harmonization is underpinned by a CDM. Additionally, BIO recommends that FDA provide a list of potential CDMs other than the ones listed in this draft guidance.

7. Validation Considerations

- BIO recommends that the draft guidance be clear and consistent in referencing that the need for and extent of validation required for a given study will depend on the specific context at-hand. Variables of importance for validation could be discussed and agreed to with FDA before the study begins. Initial discussions with FDA could include feasibility and validation study plans. BIO also suggests that circumstances when validation cannot be performed or where there is an acceptable alternative be discussed in the draft guidance.
- The draft guidance emphasizes the Agency's views on the need for validation studies for populations, exposures, covariates, effect modifiers, and outcomes. While BIO agrees with the importance of validation to ensure adequate internal validity of RWE studies, the criteria for what needs to be validated as described in the draft guidance may only be feasible in very rare circumstances (i.e. individual level validation). Baseline case report forms (CRFs) within prospective RCTs often rely on patient self-report or subjective physician judgment. For example, a patient reported history of pneumonia at baseline in a prospective RCT would not require clinical validation, review of notes, or adjudication of the clinical concept. It seems unreasonable to then require additional criteria in an EHR. In fact, in some cases, the EHR diagnosis will likely be more precise than a patient recall or physician assessment in an RCT. BIO recommends that FDA carefully reconsider guidance to validate every single variable and outcome at the individual patient level within a given study using EHR or claims data. BIO also requests

that the Agency consider providing additional guidance on how outcomes that are used in routine practice and are often the most valuable to the patient and HCP can be used in studies that incorporate EHR and claims data.

- EHR databases that use technology-enabled abstraction (e.g., when data providers use chart review at scale to create enhanced EHR databases) 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. BIO recommends that the Agency clarify if variables abstracted this way are considered 'validated' via complete verification, provided that sponsors provide descriptions of the medical record review processes, kappa statistics, etc.

8. <u>Case Examples</u>

- The draft guidance includes some conceptual examples to illustrate considerations/trade-offs, such as the example for neural tube defects in infants. However, it would be helpful to include some additional case studies of how EHRs and claims data were successfully and unsuccessfully used for efficacy studies (or if confidentiality is an issue, theoretical cases for such examples). While the draft guidance provides useful information on selection of data sources, etc., it does not provide recommendations on how submitted data from EHRs and claims data have been used by the FDA in regulatory decision making. Hence, BIO recommends that the Agency consider including more examples of RWD being used for regulatory decision-making, including rationales, in the guidance. BIO also recommends that the Agency consider development of a database or library of validation examples that would transparently guide sponsors with respect to accepted validation approaches and increase understanding in the research community of important trends by disease area or data type (i.e., EHR versus claims).

Sincerely,

/s/
Camelia Thompson, Ph.D.
Senior Director, Science and Regulatory Affairs
Biotechnology Innovation Organization

SPECIFIC COMMENTS

SECTION	ISSUE	PROPOSED CHANGE
I. INTRODUCTION AND SCOPE		
Line 32, Footnote 7	The footnote states, "For the purposes of this draft 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). Through the draft guidance, FDA uses the terms clinical studies, studies, and study interchangeably." Interventional studies may use EHR or medical claims data in different settings other than external control arm, for example, in randomized pragmatic trials, hybrid trials and decentralized trials. Consistency throughout the text regarding the use of either clinical study/studies or just study/studies throughout the document is suggested.	BIO suggests the following edit: For the purposes of this draft 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**, randomized pragmatic trials, hybrid trials and decentralized trials**) 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). Through the draft guidance, FDA uses the terms clinical studies, studies, and study interchangeably. BIO recommends that the Agency is consistent and provides clear definitions study types, including RWE studies with clinical trials, to limit confusion.
Lines 21-24 and Lines 30-32	The draft guidance states, "Pursuant to this section, FDA created a framework for a program to evaluate the potential use of real-world evidence (RWE) to help support the approval of a new indication for a drug already approved under section 505(c) of the FD&C Act or to help to support or satisfy postapproval study requirements (RWE Program)."	BIO recommends that FDA clarify whether the scope of the draft guidance includes clinical studies using EHRs or medical claims data to support new drug applications (NDA) and also address the scenario of RWE supporting an NDA as pivotal data.

SECTION	ISSUE	PROPOSED CHANGE
	The draft guidance states, "This draft guidance is intended to provide sponsors, researchers, and other interested stakeholders with considerations when proposing to use ***electronic health records*** (EHRs) or ***medical claims data*** in clinical studies to support a regulatory decision on effectiveness or safety." Lines 21-24 seem to suggest that only studies intended to support new indications (i.e. sNDA) for approved drugs or studies to satisfy postapproval study requirements are in scope. However, EHRs and medical claims data on the current standard of care could play a role in studies supporting a new drug application, for example, used to create an external control arm for a pivotal single-arm trial. Excluding such studies from the scope of this draft guidance may be a missed opportunity.	
Line 32 and Lines 97-98	The draft guidance states, "…clinical studies to support a regulatory decision on effectiveness or safety." The draft guidance states, "For all studies using EHRs or medical claims data that will be submitted to FDA to support a **regulatory decision**, sponsors should submit protocols and statistical analysis plans…" As real-world data (RWD) can support regulatory decisions in many contexts, e.g. rationale for target patient population (trial design discussions/unmet need), it would be helpful for the Agency to define all	BIO recommends that the Agency reference specific types of regulatory decisions this draft guidance would apply to.

SECTION	ISSUE	PROPOSED CHANGE
	the specific types of regulatory decisions this draft guidance would apply to.	
Lines 36-37	The draft guidance states, "RWD are data relating to patient health status or the delivery of health care routinely collected from a variety of sources." RWD is defined as data "routinely collected". This does not seem to apply to e.g., non-interventional observational studies with prospectively planned data collection, although such studies are in scope (footnote 7). The prospectively planned data collection will include many routine data but may extend beyond routine depending on the question to be answered.	BIO recommends that the Agency clarify whether the term 'routinely collected' in the definition of RWD means secondary use of claims and EHR and excludes data collected from claims or EHR from set visits per a prospectively planned study. BIO also recommends that the Agency clarify if pragmatic trials, non-interventional observational studies with prospectively planned data collection and other innovative hybrid designs that augment prospective data collection with data collection from electronic healthcare data would fall in the scope of the draft guidance if the visits are prospectively planned and not considered "routine".
Lines 54-56	The draft guidance states, "Selection of data sources that appropriately address the study question and sufficiently characterize study populations, exposure(s), outcome(s) of interest, and key ***covariates***" BIO notes that it is not always straightforward to precisely characterize the question that is to be addressed with a clinical study using EHRs or medical claims data to make inferences about the causal effect of a drug.	BIO recommends that the Agency clarify that approaches such as the target trial framework (Hernan MA, Robins JM. Am J Epidemiology 2016; 183:758) and estimand thinking process (as described in the ICH E9 Addendum) can be useful for formalizing the 'study question' when this concerns the causal effect of a drug on an efficacy or safety variable. A combination of the target trial and ICH E9(R1) estimand frameworks can be useful here to enable a precise definition of the causal estimand(s) of interest. Defining key estimand(s) can bring clarity to study design and analysis, helping to ensure all relevant data are considered for collection. Limitations of different data sources are thus more transparent, and an assessment can be made as to whether the data are sufficient to address the study question. Hence, BIO suggests that FDA expand the current draft guidance to highlight the role of estimands and the target trial framework for informing data selection and study design.

SECTION	ISSUE	PROPOSED CHANGE
		Similarly, lines 507-508 and line 991 refer to other draft guidance documents on RWE study design. BIO would welcome a more detailed discussion in these future draft guidance documents of how to apply the estimand thinking process to clinical studies using EHRs and claims data, and how to align the analysis strategy with the causal estimand.
Lines 54-62	Lines 26-29 (page 1) are quite general, yet line 54-62 in (page 2) as well as repeated use of "study hypothesis", "study question", "confounder", validation of exposure and outcome throughout the document implies that the draft guidance is primarily concerned with use of RWD to answer a causal inference-type question of whether a therapy causes an outcome. Because the draft guidance recommends pre-specification of protocols and SAP, interaction with regulators, and validation of multiple design attributes (see lines 97-104) clarifying the scope is essential.	BIO recommends that the Agency clarify the range of 'study questions' in scope of this draft guidance. Specifically, BIO recommends the Agency clarify if only questions targeting causal estimands in scope (e.g. as would be targeted by comparative safety and effectiveness studies), or does the scope also include questions concerning non-causal estimands such as statistics characterizing disease epidemiology, natural history, the benefit-risk landscape or drug utilization patterns.
Page 2, Line 60	The draft guidance states, "Data provenance and quality during data accrual, data curation, and into the final study specific dataset" It is unclear if this topic is restricted to documentation of study-specific data management or should provide documentation and justification of the whole data generation process.	BIO recommends the following edit: ~~**Data provenance and quality during data accrual, data curation, and into the final study specific dataset**~~ **Evaluation of data reliability (data accrual, completeness traceability) from provenance to final study report.**
Footnote 7, page 6	The footnote states, "For the purposes of this draft guidance, the term *clinical studies*… This footnote defines the clinical studies in scope. Observational studies are referred to as in scope.	BIO recommends that the Agency clarify the term "Observational study".

SECTION	ISSUE	PROPOSED CHANGE
	However, prospectively planned data collection per e.g. e-CRF is not discussed as a data source	
Lines 66-69 and 397-402	The draft guidance states, "For the purposes of this draft guidance, the term *reliability* includes data ***accuracy, completeness***, provenance, and ***traceability***. The term *relevance* includes the availability of key ***data elements*** (exposure, outcomes, covariates) and sufficient numbers of representative patients for the study." The terms "reliability" and "relevance", which were discussed in the FDA's Framework for RWE, are introduced in this section. However, "reliability" has only 2 mentions in the rest of the draft guidance aside from this defining statement. EHR accuracy, completeness, provenance, and traceability are predominantly site/vendor owned.	BIO recommends that FDA consider more clearly describing where "reliability" fits into the body of the document and FDA's recommendations around data quality. BIO recommends the following edits: "For the purposes of this draft guidance, the term *reliability* includes data ***accuracy, completeness***, provenance, and ***traceability***. The term *relevance* includes the availability of key ***data elements*** (exposure, outcomes, covariates) and sufficient numbers of representative patients for the study **and sufficient follow-up time**." BIO recommends that FDA provide additional draft guidance to EHR/medical claims owners and vendors that broker data to ensure accuracy and completeness. The research community would benefit from an understanding of what level of traceability would be considered minimally viable for acceptance.
II. BACKGROUND		
III. GENERAL CONSIDERATIONS		
Lines 97-100	The draft guidance states, "For all studies using EHRs or medical claims data that will be submitted to FDA to support a regulatory decision, sponsors should submit protocols and statistical analysis plans before conducting the study. Sponsors seeking FDA input before conducting the study should request comments or a meeting to discuss the study with the relevant FDA review division."	BIO suggests the following edits: "For all studies using EHRs or medical claims data that will be submitted to FDA to support a regulatory decision **for adequate and well controlled (A&WC) evidence**, sponsors should submit protocols and statistical analysis plans before conducting the study. Sponsors **should seek** ~~**seeking**~~ FDA input before conducting the study **and** should request comments **via IND submission** or a meeting **(i.e., Type C, Type D, RWE Pilot Program)** to

SECTION	ISSUE	PROPOSED CHANGE
	The draft guidance suggests that all protocols and SAPs should be discussed with or sent to FDA prior to being conducted. It is unclear what the Agency means by "before conducting the study". The last statement makes it seem like seeking FDA input before conducting the study is optional, but in later sections, there are comments that sponsors should discuss definitions before study conduct. This sentence should reflect that FDA input is necessary, rather than optional. Other discussions in the draft guidance should be consistent. Comparative safety and effectiveness studies typically proceeds in multiple iterative stages with refinement of the design and questions during a fitness-for-purpose and feasibility period followed by finalizing the SAP and analytical period. The earlier period sometimes includes a few analytical results to assess cohort size and study power. It is unclear if Sponsors should interact with the Agency prior to any analysis or prior to finalizing the analyses to address the primary hypotheses or both.	discuss the study with the relevant FDA review division **and appropriate review staff**. " BIO recommends the Agency clarify the meaning of "before conducting the study". It would be helpful for FDA to elaborate on the appropriate mechanism to gain timely and efficient feedback from appropriate FDA review staff, such as a Type C meeting and/or considering the new RWE pilot program under PDUFA VII and Type D meeting option. Lack of clarity on the specific way(s) to obtain feedback may lead to inconsistencies in how sponsors approach FDA, and could lead to minimal interaction. Similarly, BIO recommends that the Agency clarify if the draft guidance recommends interactions with the Agency prior to any analysis or prior to finalizing the analyses to address the primary hypotheses or both. BIO also recommends that the essential pre-defined elements of study design, analysis, conduct and reporting be part of the scope of the FDA interaction to ensure that it is clear that FDA input should be sought on these specific points. It would also be helpful for FDA to clarify whether the suggestion is for studies intended to be registrational (i.e., providing substantial, adequate and well-controlled (A&WC) evidence and when EHR and medical claims data are used to provide such evidence for primary and key secondary endpoints). Further, we recommend that FDA provide examples within this section or in an appendix of types of studies using EHR/claims that support a regulatory decision and the suggested interaction with FDA.

SECTION	ISSUE	PROPOSED CHANGE
Lines 110-125	The draft guidance states, "This draft guidance addresses issues that are essential to determining the reliability and relevance of the data and that should be addressed in the protocol, including:…" The FDA should clarify if the Sponsor is working with a data vendor to collect RWD, if unpublished statements from the data vendor and/or must the data vendor have published works that can be used to address issues around "(1) the appropriateness and potential limitations of the data source for the study questions and to support key study elements" and "(4) Quality assurance and quality control (QA/QC) procedures for data accrual, curation, and transformation into the final study-specific dataset.	BIO suggests the following edit: "This draft guidance addresses issues that are essential to determining the reliability and relevance of the data and that should be addressed in the protocol **(e.g., with peer-reviewed literature)**, including:…"
Lines 124-125	The draft guidance states, "Quality assurance and quality control (QA/QC) procedures for data accrual, curation, and transformation into the final study-specific dataset."	BIO recommends the following edits: "Quality assurance and quality control (QA/QC) procedures for data accrual, curation, **data integration/linking and data transfer/movement** and transformation into the final study-specific dataset."
IV. DATA SOURCES		
Entire Section	Based on the information provided in this section, we suggest that FDA put together a checklist for assessing each data source that would be included as part of the protocol. It may be helpful here to describe the information that is available from EHRs and claims data. For example, EHRs may have information on inpatient medication, lab results, vital signs, behavioral risk factors, family history data, etc. that are not typically available in claims. Although actual information available in a specific EHR system will depend on health care	BIO recommends that the Agency consider developing a checklist for assessing each data source that would be included as part of the protocol. BIO also suggests that the Agency consider describing the information that is available from EHRs and claims data.

SECTION	ISSUE	PROPOSED CHANGE
	systems and users, there are general differences between claims and EHRs that can be highlighted.	
Lines 135-136	The draft guidance states, "Each data source should be evaluated to determine whether the available information is appropriate for addressing a specific study hypothesis." It is unclear if the evaluation of each data source should highlight how the data source was selected vs. other data sources that were rejected for delineated reasons.	BIO suggests the following edit: **An analysis of** each data source should be **conducted to evaluate** ~~**evaluated to determine**~~ whether the available information is appropriate for addressing a specific study hypothesis**, and should be made available to the FDA upon request.** BIO recommends that the Agency clarify if the evaluation of each data source should highlight how the selected fit-for-purpose data source was selected vs. other data sources that were rejected for delineated reasons.
Lines 141-143	The FDA draft guidance states, "The purpose of medical claims data is to support payment for care; claims may not accurately reflect a particular disease, or a patient may have a particular disease or condition that is not reflected in claims data." As presented, this description regarding claims might be too simple and not fully capture the nuances of limitations.	BIO recommends the following edits: ~~**The purpose of medical claims data is to support payment for care**~~; **C**laims may not accurately reflect a particular disease **and the comprehensive management/treatment (e.g., care not reimbursed by insurance),** or a patient may have a particular disease or condition that is not reflected **or well reflected (e.g., not unique codes)** in claims data. BIO recommends that the Agency considering noting that that the claim for the treatment with a drug may not reflect the actual use of the drug.
Lines 145-150	The draft guidance states, "EHR data are generated for use in clinical care and may also serve as a basis for billing and for auditing of practice quality measures. Data recorded in an EHR system depend on each health care system's practices for patient care and the clinical practices of its providers. In addition, data collection is limited to the data captured within an EHR system or network, and may not represent comprehensive care (e.g., care obtained outside of the health care system)."	BIO recommends the following edits: EHR data are generated for use in clinical care and may also serve as a basis for billing and for auditing of practice quality measures. Data recorded in an EHR system depend on each health care system's practices for patient care and the clinical practices of its providers **as well as the providers' documentation habits (e.g., level of detail captured)**. In addition, data collection is limited to the data captured within an EHR system or network, and may not represent comprehensive care (e.g., care obtained **in different facilities in the same or** outside of the health care system).

SECTION	ISSUE	PROPOSED CHANGE
	In the description of EHRs, there is the suggestion that such systems are prone to data gaps in instances where care is sought outside the health care system. Because there is sometimes confusion between the terms EHR and EMR and they tend to be used interchangeably, it may be useful to also define what is meant by EMR databases at the outset. The limitations presented regarding health care practices does not reflect actual documentation habits of the provider. Additional clarification would inform on the limitations of the data assessed based on details captured. The same health care system may use different EHR in different facilities. In addition, data collection is limited to the data captured within an EHR system or network. EHR data can be subject to similar issues as claims, i.e., wrongly classified, incorrectly specified, or missing.	**Similar to claims, EHR data may not accurately reflect presence, specificity, or severity of a particular disease.** BIO recommends that the Agency consider defining EMR in the glossary to differentiate between an EHR and EMR database and state that the draft guidance focuses on EHR data. BIO recommends that the Agency clarify if this description of EHR includes abstracted data.
Lines 152 to 156	The draft guidance states, "For prospective clinical studies proposing to use EHRs, it may be possible to modify the EHR system for the purpose of collecting additional patient data during routine care through an add-on module to the EHR system. However, given the limited ability to add modules to collect extensive additional information, EHR-based data collection may still not be comprehensive." Extensive data capture via EHR might be limited, but using new digital technology, additional data needed for prospective research can be supplemented .	BIO recommends the following edit: "For prospective clinical studies proposing to use EHRs, it may be possible to modify the EHR system for the purpose of collecting additional patient data during routine care through an add-on module to the EHR system. However, given the limited ability to add modules to collect extensive additional information, EHR-based data collection may still not be comprehensive. **While extensive data capture via EHR may be limited, the use of digital technology may provide additional data needed for prospective clinical studies.**"

SECTION	ISSUE	PROPOSED CHANGE
	There is no detail on examples of acceptable add-on module to the EHR system. There is no draft guidance or expectation of using the Estimand framework in the context of RWE.	BIO recommends that the Agency provide examples of acceptable add-on modules to EHR systems. BIO also recommends including a reference as to where the use of Estimands, as discussed in ICH draft guidance E(R1), may be relevant in the context of RWE.
Lines 158-161	The draft guidance states, "The historical experience with and use of the selected data source for research purposes should be described in the protocol. This description should include how well the selected data source has been shown to capture study elements (e.g., inclusion and exclusion criteria, exposures, outcomes, key covariates) and how the data can be validated for a particular research activity." We note that not all data needs to be validated. Because the term 'validated' has specific downstream implications that additional validation studies need to be done just on the specific study element, we recommend alternative wording. Additional detail is needed in terms of "*The historical experience with and use of the selected data source for research purposes should be described in the protocol*." to capture the experience of the researcher. Even if the researchers have experience of the selected data source, experience in a different therapeutic area (TA) might not be as helpful in a new TA for which there is less experience (e.g., experience in acute but limited experience in chronic) because providers in different TA might document differently or the billing patterns might be different. An investigation of whether the current EHR or claims carried that information needed in a new TA will be necessary.	BIO suggests the following edits: The historical experience with**, including relevant experience in the specific therapeutic area proposed,** and use of the selected data source for research purposes should be described in the protocol. This description should include how well the selected data source has been shown to capture study elements (e.g., inclusion and exclusion criteria, exposures, outcomes, key covariates **and intercurrent events**) and how the data can be ~~**validated**~~ **suitable** for a particular research activity.

SECTION	ISSUE	PROPOSED CHANGE
	In order to accurately estimate the primary estimand, the data source should also capture intercurrent events which could preclude observation of the outcome variable or affect its interpretation (e.g. use of rescue medication, treatment switching or treatment discontinuation, etc). For pragmatic reasons, one could prioritize the 1-2 intercurrent events which are anticipated to occur with highest frequency and /or have the largest impact on study results.	
A. Relevance of Data Source		
Lines 166-169	The draft guidance states, "Patients in different types of commercial or government health care payment programs can differ in a range of characteristics, such as age, socioeconomic status, health conditions, risk factors, and other potential ***confounders***." Patient characteristics listed are not always confounders in every study. Some variables might be associated with outcome only, or exposure only. Covariates would be a better term here than confounders since covariates include both confounders and effect modifiers as defined in the glossary.	BIO suggests the following edit: Patients in different types of commercial or government health care payment programs can differ in a range of characteristics, such as age, socioeconomic status, health conditions, risk factors, and other potential **covariates** ~~***confounders***~~."
Lines 169-173	The draft guidance states, "Various factors in health care systems and insurance programs, such as medication tiering (e.g., first-line, second-line), formulary decisions…" Use of the term medication tiering (e.g., first-line, second-line) is confusing since it refers to co-payment system (examples should be Tiers 1-4, not for first-line,	BIO suggest the following edit: Various factors in health care systems and insurance program, such as **patient out-of-pocket payment** ~~**medication tiering (e.g., first-line, second-line)**~~, formulary decisions…"

SECTION	ISSUE	PROPOSED CHANGE
	second-line since first-line, second-line can be confused with the line of therapy.	
Line 178	The draft guidance states, "The reason for selecting the particular data sources to address the specific hypotheses." Many RWE studies are not necessarily hypothesis driven.	BIO recommends that the Agency clarify if this draft guidance document only applies to hypothesis driven studies.
Lines 180-185	The draft guidance states, "2. Background information about the health care system, including (if available) any specified method of diagnosis and preferred treatments for the disease of interest, and the degree to which such information is collected and validated in the proposed data sources. 3. A description of prescribing and use practices in the health care system (if available), including for approved indications, formulations, and doses." EHR and Claims data usually cover many health care systems. It is not practical and may not even be possible to provide the requested background information about prescribing practices and preferred treatments.	BIO recommends that the Agency clarify what validation means in this statement.
Lines 184-185	The draft guidance states, "A description of prescribing and use practices in the health care system (if available), including for approved indications, formulations, and doses." There is a typo, i.e., an additional "for".	BIO suggests the following edit: A description of prescribing and use practices in the health care system (if available), including ~~for~~ approved indications, formulations, and doses
Lines 187-188	The draft guidance states, "For non-U.S. data sources, FDA recommends providing an explanation of how all of these factors might affect the generalizability of the study results to the U.S. population."	BIO recommends that the Agency clarify whether evaluation of generalizability applies to all data sources, including cohorts derived from RWD, regardless of whether they are US based with additional considerations to non-US based databases.

SECTION	ISSUE	PROPOSED CHANGE
	Some US databases include a fraction of the US population. Thus, their generalizability to the whole US population can be called into question. Furthermore, study design strategies to minimize bias (e.g., incident use, active comparator, minimum look back for confounders) can further limit generalizability of the results.	BIO recommends that the Agency clarify to what extent would factors impacting generalizability to a US population influence the FDA's acceptance of non-US RWD and if there are any factors that are of greatest concern to generalizability (and thus should be listed) from the Agency's perspective. BIO also recommends that the Agency highlight specific circumstances in which they may not accept a non-US data source (e.g., when race may be an important factor for a disease).
B. *Data Capture: General Discussion*		
Lines 196-199	The draft guidance states, "Sponsors should demonstrate that each data source contains the detail and completeness needed to capture the study populations, exposures, key covariates, outcomes of interest, and other important parameters (e.g., timing of exposure, timing of outcome) that are relevant to the study question and design." If the sponsor is working with a data vendor to collect RWD on a specific population of interest, access to data completeness outside of the selected population may not be easily obtained. The draft guidance needs to provide a definition and some examples for "key covariates".	BIO recommends that the Agency provide clarity on readily accepted standards/methods of what completeness is needed and ways in which to "demonstrate" completeness. BIO also recommends that the Agency provide a definition and some examples for "key covariates".
Entire section	This section includes details on the data specific information that should be shared. The section describes the need for submitting information such as "length of follow-up to ascertain outcomes" and "distribution of length of follow-up for patients in the data sources".	BIO recommends that the Agency clarify whether these details are requested for the study population of interest or for the overall population included in the data source. Typically, the study specific population data are not analyzed prior or available at the time of protocol/analysis plan development.
1. *Enrollment and Comprehensive Capture of Care*		

SECTION	ISSUE	PROPOSED CHANGE
Lines 203-205	The draft guidance states, "***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." EHRs may not capture enrollment and disenrollment, but this information may be approximated by time period the patient is active in the database. Enrollment capture is an important element when working with RWD. It should be highlighted in the draft guidance that many oncology EHRs do not have enrollment information on patients and rely merely on last health care interaction as a means of determining whether a patient is still in the health care system. Gaps in coverage or care sought outside the healthcare system are not usually captured in these systems and can only be inferred by pauses in health care interactions/visits. This is an important limitation to highlight as there may be the introduction of immortal time as well as missing exposure data. Requiring specific enrollment or coverage times could bias a cohort towards healthier populations; additional	BIO suggests the following edit: ~~*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.*~~ **For claims data, enrollment/disenrollment can be assessed based on insurance enrollment and disenrollment information. For EHRs, this information may not be accurately captured and can only be approximated. Continuity of medical coverage within an EHR network should be addressed when using EHR data sources.** BIO recommends that the Agency consider mentioning that many oncology EHRs do not have enrollment data and potentially propose later in the draft guidance an approach to deal with this limitation. BIO recommends that language characterizing continuity of care should be delineated (where relevant) to the use of EHR and medical claims data separately, as well as jointly, if applicable. The draft guidance should be specific that EHR and claims cases are different and that claims are more susceptible to issues with continuity of coverage. The Agency should reference lines 588-625 as relevant to claims data and not EHR. BIO recommends that the Agency consider using the term "***Continuity of observable period***" to indicate the period that patients remain in the healthcare system.

SECTION	ISSUE	PROPOSED CHANGE
	methods of ensuring "coverage" could be considered (e.g., Requiring a certain length of potential follow-up time from an appropriate index date may also ensure patients have enough documentation ingested to support data collection and analysis). It appears that FDA may use claims data and EHR data interchangeably. In this section, sometimes the terminology (e.g. "health systems", "coverage", "enrollment", etc.) is not explicit on whether it is referring to recommendations that apply to both insurance networks and EHR networks, or only one or the other. Rationale: This section describes a recommendation on continuity of coverage that mentions both EHR and medical claims data, although the described limitation applies primarily to medical claims data and not EHR. For instance, insurance plan disenrollment would introduce such a "discontinuity" of coverage described here; however, for databases solely using EHR data, such an interruption is not necessarily observed. A related (and more broad) aspect of "continuity of care" would apply to both, in which patients seek care outside of a system included in an EHR network (e.g., specialist in a different health system) and that information is unobservable to the real world dataset. Continuity of coverage applies to claims data only. For EHR data, a more appropriate concept might be an	

SECTION	ISSUE	PROPOSED CHANGE
	indicator of activity in the database (e.g. visits, labs, etc.). The term "***continuity of coverage*** (enrollment and disenrollment)" may not be applicable when referring to EHR data sources, as enrollment and coverage cannot be determined in EHR data.	
Lines 205-207	The draft guidance states, "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." As opposed to claims, many EHR systems do not contain an enrollment file.	BIO recommends that the Agency consider adding draft guidance on how to define patient eligibility in those data sources, such as EHR systems that do not contain an enrollment file.
Entire Section	There are some attributes of open claims that FDA may want sponsors to report.	BIO recommends that FDA distinguish between closed claims sources and open claims sources in the draft guidance.
Lines 217-219	The draft guidance states, "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." It might be helpful to clarify this example. Real-world testing practices have much greater variability than what would be seen in a clinical trial, and that's to be expected. It's unclear why an outcome that is dependent on a specific frequency of laboratory tests would be chosen for a study that depends on real-world testing practices. The second example provided in the documentation does not reflect the issue of comprehensiveness of the data sources in capturing aspects of care and	BIO recommends that the Agency clarify the current example and consider including the more appropriate example of an outcome that is not captured in the medical claims data source such as laboratory results.

SECTION	ISSUE	PROPOSED CHANGE
	outcomes that are relevant to the study question. In this case, the data source may comprehensively capture the outcome, but the outcome may still rarely show up as it is not a norm for clinicians to order those tests indicative of the outcome.	
Lines 223-226	The draft guidance states that "The data sources should contain adequate numbers of patients with adequate length of follow-up…"	BIO recommends that the Agency clarify if a feasibility assessment be undertaken from various data sources to address the length of follow up available, before selecting the final data source(s).
Lines 231-235	The draft guidance states, " 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 questions, the data source may not be suitable, or the protocol should describe how this information gap will be addressed." It is unclear how information gaps should be addressed.	BIO suggests the following edits: "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 questions, the data source may not be suitable, or the protocol should describe how this information gap will be addressed **(e.g., adding additional modules to the EHR system to address the information gap or collecting the data outside the EHR system)**."
2. *Data Linkage and Synthesis*		
Entire Section	Linkages are emphasized in the document, but trade-offs with decreased sample size should be acknowledged. While draft guidance for analysis methods is not provided in this document, certain considerations for analysis using linked datasets should be taken at the design stage (e.g. assessing the noninformative linkage assumption, which states that the linkage process is conditionally independent from the outcome and covariate distributions in the analysis given the identifying information used to perform the linkage).	BIO recommends that the Agency clarify that the linkage rate is an important factor for feasibility consideration, since, if it is too low, it may lead to a sample size that is too small and not meaningful. This clarity will allow Sponsors to explore both deterministic and probabilistic options. BIO recommends that the Agency clarify that if the Sponsor is working with a data vendor to collect RWD, for which a data linkage already exists, whether unpublished statements from the data vendor on linkage quality can be provided to the Agency or must the data vendor have published works addressing the linkage quality.

SECTION	ISSUE	PROPOSED CHANGE
	When linking data sets together, variables of interest that exist in both data sources may not agree in value. Draft guidance on how to resolve disagreements would be helpful (e.g., sponsor should provide justification for selecting a particular set of values from one data source over another; assessing degree of agreement; conducting sensitivity analyses)	BIO recommends that the Agency describe data tokenization as a method for linking data. BIO recommends that the Agency clarify whether steps should be performed on linked data sets to obtain a 1:1 linkage (e.g., linear programming methods are commonly implemented after probabilistic linkage as a greedy method to obtain 1:1 linkage).
Lines 250-251 and 1125-1198	The draft guidance states, "Data linkages can be used to increase the breadth and depth of data on individual patients…" Unclear whether those terms (breadth and depth) refer to the size of the cohort and number of covariates on each subject or more specifically refers to longitudinality Patients may or may not be included in multiple health care site sources, which may be challenging to identify in a multi-source aggregate. EHR/medical claims data sources are also in a persistent state.	BIO recommends that the Agency clarify the meaning of breadth and depth of data by adding this terminology to glossary or adding a reference BIO recommends that the agency provide specific recommendations on reconciliation of such cases where patients may or may not be included in multiple health care site sources (i.e., to ensure accuracy, completeness, and non-duplication). BIO recommends that the Agency clarify if Sponsors should routinely evaluate data changes at the source to address situations where, for example, a patient could update their list of medications after an initial cut of study data.
Lines 251-255	The draft guidance states, "If the study involved establishing new data linkages between internal data sources (e.g., mother-infant linkages) or external data sources…" Reference to internal and external in the context of this text are meant to characterize the data elements, not the data sources, which could be confused with internally- or externally held data sources.	BIO recommends the following edit: "If the study involved establishing new data linkages between **data elements within the same** ~~**internal**~~ data sources (e.g., mother-infant linkages) or **data elements across different** ~~**external**~~ data sources

SECTION	ISSUE	PROPOSED CHANGE
Lines 269-272	The draft guidance states, "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." It is not clear what acceptable quality means. Many more characteristics may vary across data sources but only a few of those would have an impact on bias or accuracy of the causal inference This section discusses approaches for combining data sources. Additionally, the same patient's data may exist in multiple data sources. When combining them, duplicate records should be deleted.	BIO recommends that the Agency provide more clarity on what "acceptable quality" would be in this context or at least an example on approach. BIO recommends that the Agency clarify whether any heterogeneity should be documented on only those characteristics that are relevant for the main study hypothesis. BIO recommends that the Agency include a rationale of circumstances where it is applicable to combine data sources.
Lines 282-286	The draft guidance states, "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." Even with data sources sharing a unique patient identifier, or single data sources with many data elements all linked through a unique identifier, the existence of multiple records and duplicate records is still an issue that should be addressed in a study	BIO recommends the following edit: ~~**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.**~~ **The presence of multiple records or duplicate records is an issue even within one single data source or multiple data sources linked by a unique patient identifier. This requires curative action, which should be documented and justified in a study protocol by researchers prior to initiation of the study.** 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."

SECTION	ISSUE	PROPOSED CHANGE
	protocol. Researchers should document the type of curation performed to handle the issue of multiple records or duplicate records even when a unique identifier is present. For example, an individual in claims data may have duplicate records for the same prescription. Which record will the researcher keep? Maybe the record with the longest prescription length (based on days supply). Or an individual likely has duplicate or overlapping records in confinement data, so researchers must formulate a plan as to which record to keep. For sponsor working with 3rd party data vendors, it is recommended that the documentation can also be provided in conjunction with a 3rd party data vendor.	BIO recommends that the Agency provide an example to clarify this point.
3. *Distributed Data Networks*		
Line 329-330	The draft guidance states, "Transforming disparate database structures into a common health network with a CDM allows research across health care sties that would otherwise be more complex and costly." The FDA should clarify if they have a preferred CDM in which to receive the RWD.	BIO recommends that the Agency reference the recent draft guidance on Data Standards for Drug and Biological Product Submissions Containing Real-World Data.
Entire section	The amount of information on distributed data networks (DDN) is excessive. The data sources and integrated data used from a DDN should be assessed in the same manner as any other data, and data transformation is already covered elsewhere.	
Lines 292-294	The draft guidance states, "…often combined with the use ***Common Data Models*** (CDMs),…" Add "of" between use and Common Data Models	BIO suggests the following edit: "…often combined with the use **of** ***Common Data Models*** (CDMs),…"

SECTION	ISSUE	PROPOSED CHANGE
292-296; 323-327; 340-350; 1169-1175	A recommendation is given to use Common Data Models (CDMs) to harmonize the data sources. However, there's no draft guidance on what CDM can/could/should be used. The research community would benefit from further details on common data models (CDMs) that are already well developed, such as OMOP/ODHSI and HL7/FHIR, and an indication if FDA prefers these CDMs. It would be helpful to understand if FDA would prefer to have data mapped to these newer standards, or would they be willing to accept data in native format since all the RWD CDMS are still evolving.	BIO recommends that the Agency reference the recent draft guidance on Data Standards for Drug and Biological Product Submissions Containing Real-World Data. BIO recommends that the Agency clarify they endorse, for example, the use of non-CDISC data standards (e.g., HL7 FHIR or OHDSI OMOP CDM). Any input on what other CDMs that are acceptable not already described as part of the FDA Data Standards Catalog would eliminate sponsor assumptions, unexpected downstream technical limitations, or industry variance in provided CDMs.
4. *Computable Phenotypes*		
Lines 363-365	The draft guidance states, "The computable phenotype definition, composed of data elements and phenotype algorithm, should be described in the protocol and study report and should also be available in a computer-processable format."	BIO recommends that the Agency define what a computer-processable format is and clarify if this means that the code used to compute the phenotypes should be available and reproducible. For example, if the algorithm is based on a trained machine learning model, FDA should clarify if both the model and the training code should be made available and documented.
Lines 365-366	The draft guidance states, "Clinical validation of the computable phenotype definition should be described in the protocol and study report." It is unclear what should be described int the protocol and study report.	BIO recommends that the Agency clarify what they expect sponsors to 'describe in the protocol and study report' with respect to clinical validation.
Entire section	The term "computable phenotype" isn't useful in the context of this draft guidance. A computable phenotype is essentially an operational definition for the target patient population (i.e., a code-based electronic algorithm using structured data elements).	BIO recommends using "operational definition for the target patient population" for consistency [with later sections of the draft guidance that address operational definitions for exposures, outcomes, and covariates].
5. *Unstructured Data*		

SECTION	ISSUE	PROPOSED CHANGE
Line 383	The draft guidance notes that "FDA does not endorse any specific AI technology."	BIO recommends that the Agency clarify as to whether FDA would consider AI extracted endpoints from text data sources.
Line 388-393	The draft guidance states, "If the protocol proposes to use AI or other derivation methods…" It is important that researchers transparently present the details of any AI or other derivative methods they plan to use in the study. Often this would be lengthy and highly complex on its own. If the protocol proposes to use AI or other derivation methods, the protocol should specify the assumptions and parameters of the computer algorithms used, the data source from which the information was used to build the algorithm, whether the algorithm was supervised (i.e., using input and review by experts) or unsupervised, and the metrics associated with validation of the methods. Relevant impacts on data quality should be documented in the protocol and analysis plan. It needs to be clarified that this is relating directly to the use of AI or other extraction methods and not human abstraction. Any assessment of algorithm performance in relevant subgroups (i.e., to ensure that algorithms do not introduce or perpetuate measurement bias) should be reported. In several cases, the draft Draft guidance mentions that certain source data and algorithms should be provided in the protocol. This is most relevant as it pertains to AI-	BIO suggests the following edit: "If the protocol proposes to use AI or other derivation methods…Relevant impacts on data quality **from use of AI or other derivation methods** should be documented in the protocol and analysis plan." BIO recommends that the Agency consider adding a statement that the details may be provided as an appendix/supplement to the protocol or may refer to publication(s) as reference (if applicable).

SECTION	ISSUE	PROPOSED CHANGE
	based extraction tools for unstructured data. These algorithms are proprietary in nature and the expectations outlined in the draft guidance would necessitate data providers sharing IP with drug sponsors. An alternative solution would be to provide descriptions of the source data and algorithms and then to have more detailed data available for FDA's review upon request or upon inspection.	
C. *Information Content and Missing Data: General Considerations*		
Entire Section	It would be helpful for FDA to comment on if missing data presents, if FDA's preference is for sponsors to collect additional information on the missing data, or assess the impact of missing data on study results via analytical approaches.	BIO suggests that FDA should clearly state their preferred approach to address complete or partial missing data in the EMR/claim database within Section C of this draft guidance document
Line 395	The draft guidance states, "C. Missing Data: General Considerations" The title is inconsistent with what is in the Table of Contents (C. Information Content and Missing Data: General Considerations.).	BIO recommends changing the title of this section for the reasons stated in the text. Missing data is an often-misunderstood issue and using it as a title may perpetuate the confusion. Suggested alternatives: "Unavailable Data" or "Gaps in Data"
Lines 401-402	The draft guidance states, "It is important to distinguish between these two cases and understand the reasons why information in present or absent in EHRs and medical claims." An important distinction between different types of missing data is provided in the draft draft guidance. We believe that additional discussion could be provided in the final draft guidance.	BIO recommends that the final draft guidance discuss missing data that result from differences that relate to variations in assessment that may exist between routinely collected data (e.g., EHRs) and clinical trials (related to the second category described in the draft guidance). An example would be performance status assessments such as ECOG scores which differ in terms of completeness in the clinical practice settings versus clinical trials.
Lines 418-422	The draft guidance states, "The protocol and the statistical analysis plan should be developed…"	BIO recommends that the Agency emphasize that the estimand must be defined before missing data can be discussed and an analysis plan specified.

SECTION	ISSUE	PROPOSED CHANGE
	Both the ICH E9 addendum and the FDA commissioned NRC report "The Prevention and Treatment of Missing Data in Clinical Trials" emphasize that first the estimand has to be defined before missing data can be discussed. The current draft guidance should be aligned with the text of these two references.	BIO recommends that the Agency clarify if they further recommend application of missing data statistical methods once descriptions of the reasons for missing data and missing data mechanisms have been addressed in the statistical analysis plan and protocol; Or, will description of the problem of missing data suffice.
Line 420 -422	The draft guidance states, "Assumptions regarding the missing data (e.g., missing at random, missing not at random) underlying the statistical analysis for study end points and important covariates should be supported and the implications of missing data considered." This statement is not clear as to whether FDA prefers the use of appropriate sensitivity analysis to quantify the impact of missing data on the study results. The end points should be endpoints. The use of descriptive analyses to characterize both the presence and absence of data (missing or absent) has limited feasibility. It is difficult to determine if missing data in the EHR were intentional or not.	BIO suggests the following edit: "Assumptions regarding the missing data (e.g., missing at random, missing not at random) underlying the statistical analysis for underlying the statistic analysis for study ~~end points~~ **endpoints** and important covariates should be supported and the implications of missing data considered." BIO recommends that the Agency clearly state their preferred approach to assess the implication of missing data.
D. *Validation: General Considerations*		
Lines 455-461	The draft guidance states, "To understand how potential misclassification of a variable of interest (e.g., exposure, outcome, covariate) might impact the measure of association and the interpretation of results, sponsors should consider: (1) the degree of misclassification; (2) differential versus non-differential misclassification (e.g., differential misclassification of outcome by exposure); (3) dependent versus independent misclassification (e.g., correlated misclassifications of exposure and outcome when both	BIO recommends that the final draft guidance should acknowledge that different magnitudes of confounding may necessitate different levels of rigor with respect to validation. BIO suggests the following edit: "... differential misclassification of outcome by ~~exposure~~ **treatment group** …" BIO recommends including a reference to the later section (V.D.3) where these are described in more detail.

SECTION	ISSUE	PROPOSED CHANGE
	are self-reported in the same survey); and (4) the direction toward which the association between exposure and outcome might be biased." This section seems to be missing a key parameter: the association between the confounder and the exposure, and the confounder and the outcome. If either association is small, then validation may not be as critical. The word "exposure" may mean much more than "treatment groups", i.e., dose, frequency, route, etc. Here "exposure" may be better replaced by "treatment group". Differential vs non-differential misclassifications and dependent vs non-dependent misclassifications are introduced here for the first time but without context.	
428-431	The draft guidance stipulates that "A conceptual definition should reflect current medical and scientific thinking regarding the variable of interest, such as: (1) clinical criteria to define a condition for population selection or as an outcome of interest or a covariate; or (2) measurement of drug intake to define an exposure of interest."	BIO recommends the Agency develop an algorithm to define the specific condition or drug exposure.
Lines 433-439	The draft guidance states, "An operational definition should be developed based on the conceptual definition to extract the most complete and accurate data from the data source. In many studies using EHR or medical claims data, the operational definition will be a code-based electronic algorithm using structured data elements." It is unclear who would develop the operational	BIO recommends that the Agency clarify if data providers would need to provide the code to sponsors and/or FDA and if written descriptions would be adequate.

SECTION	ISSUE	PROPOSED CHANGE
	definition and if this would need to be provided to the Agency. IP concerns may also exist here.	
Lines 441-461	The draft guidance states, "Because operational definitions are usually imperfect and cannot accurately classify the variable of interest for every subject, a resulting ***misclassification*** can lead to false positives and false negatives (Table 1) and may bias the association between exposure and outcome in a certain direction and degree. Although complete verification[10] of a variable of interest minimizes misclassification and maximizes study internal validity, understanding the implications of potential misclassification for study internal validity and study inference is the key step in determining what variables of interest might require validation and to what extent..." FDA seems to be proposing metrics for misclassification. Attempting to mitigate all potential for outcome misclassification by suggesting sponsors conduct a quantitative bias analysis as a sensitivity analysis to show how outcome misclassification may impact study results. Determining misclassification has limited feasibility. The current discussion of the consequences of measurement error focuses on misclassification, and thus seems to implicitly assume that the variables of interest will be binary. However, outcome variables may be continuous, count or time-to-event etc. A more general discussion of how the impact of measurement	BIO recommends that the Agency expand the current discussion to include more general comments on how measurement error (and the performance of operational definitions) can be assessed for variables which are not binary (such as time-to-event and continuous variables). An example could be how to define severity of disease.

SECTION	ISSUE	PROPOSED CHANGE
	error can be assessed would therefore be helpful.	
Footnote 10, p. 11	The draft guidance states, "10. For the purposes of this draft 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.)"	BIO recommends that the Agency clarify if this means that abstraction (medical record review) is equivalent to "complete verification". Other language in the draft guidance seemed to indicate that medical record review may need additional verification by data linkage to other sources.
Line 444-445 **Line 463** **Lines 821-822** **Lines 888-889**	General Comment: Complete verification alone does not create high quality data. Trial data monitoring has shifted away from 100% Source Data Verification to more risk-based approaches because 100% SDV isn't particularly effective. Suggest removing the statement about complete verification being the most rigorous or adding content on risk-based approaches. The data field has moved to risk-based approaches to data verification, and 100% SDV has been shown to not impact data from trials (cite the Tufts study).	BIO recommends the following edits: "Although **risk-based approaches to source data verification** ~~**complete verification**~~ of a variable of interest minimizes misclassification and maximizes study internal validity,...." "Although **risk-based** ~~**complete**~~ verification of a study variable is considered the most rigorous approach,...." "Although **risk-based** ~~**complete**~~ verification of the outcome variable is considered the most rigorous approach,...." "In scenarios where **risk-based** ~~**complete**~~ verification of the outcome variable for each study subject is infeasible,...."
Lines 466-469	The draft guidance states, "Based on the performance measures described in Table 1, sponsors should consider whether validating the variable to a greater extent (e.g., all positives classified by the operational definition) is necessary and discuss with the relevant review division."	BIO recommends that the final draft guidance provide citations of valid algorithms or use of datasets where data are consistently abstracted may be sufficient.

SECTION	ISSUE	PROPOSED CHANGE
	The draft guidance states that validation studies are needed.	
Lines 471-480	Several general considerations are provided for data validation. However, considerations are not discussed for situations where a hybrid clinical/RWD approach is being used (such as with a single arm trial with an external RWD control).	BIO recommends including any particular considerations that may be different in validating RWD that are being used in conjunction or in comparison to clinical trial data.
Lines 479-480	The draft guidance states, "The quality of prior studies used to establish ***sensitivity, specificity***, and predictive values should always be evaluated." FDA recommends evaluating quality of prior studies that evaluated performance of an operational definition.	BIO recommends that the Agency provide clarity on what criteria sponsors should use to assess quality.
Lines 482-490	The draft guidance states "*The protocol should include a detailed description of the planned validation*…." with subsequent lines referring to *"….justification for the choice of a reference standard, validation approach, methods, processes, and sampling strategy (if applicable)*." This draft draft guidance is unclear on the intention of "planned validation". The validation of the operational definitions should be part of the feasibility step before the final protocol & analysis plan is filed (as done in the RCT Duplicate project).	BIO recommends that the Agency clarify the intention of "planned validation" and whether it applies to all variables (exposure/outcome/all covariates), or to selected key variables only (exposure/outcome/selected key covariates). BIO recommends that the Agency clarify what the design of a study intended to assess the performance of an operational definition should look like. And, that draft guidance might be added that it is advisable to validate the operational definitions of key outcome, exposure and confounder variables before conducting the study. BIO recommends that the Agency provides more clarity (including references to relevant to literature) on how to design validation study including: criteria for setting the sample size; what performance metrics can be used if the outcome, exposure or covariate is not a binary variable; and what 'good' looks like in terms of values of the sensitivity, specificity, PPV and NPV metrics mentioned for a binary variable.

SECTION	ISSUE	PROPOSED CHANGE
		BIO recommends that the Agency provides more clarity on if 'success' thresholds for these performance metrics would depend on the type of study question that is to be addressed using the RWD (i.e. better performance required for comparative effectiveness studies vs exploratory studies).
V. STUDY DESIGN ELEMENTS		
Line 500	The title of this section, "Study Design Elements", may potentially be confusing to researchers about its intent, when considering forthcoming draft guidances that will be focusing on study design choices, which this draft guidance does not discuss. Indeed, this section appears to be primarily focused on study variable definitions and how those are defined/validated in the RWD source.	BIO recommends the Agency consider an alternative title/nomenclature, such as "Defining Study Variables" or something similar.
Line 503-504	The draft guidance states, "The study questions of interest should be established first, and then the data source and study design most appropriate for addressing these questions should be determined." It's made clear in in the document that it's intended to have other RWE draft guidances focused on study design and analysis, but it could be beneficial to cross-reference E9 (R1)	BIO recommends that the Agency make reference to the estimand draft guidance E9 (R1).
A. *Definition of Time Periods*		
Entire section	Immortal time bias is an important consideration sponsors should address.	BIO suggests adding that sponsors should consider presence of immortal time during follow-up and describe how they dealt with this potential bias, potential implications on effect measures, etc.
B. *Selection of Study Population*		
Lines 545-550	The draft guidance states, "Key variables used to select the study population should be validated. For example, to assess the drug effect in patients with	BIO suggests the following changes: Key variables used to select the study population should be validated. For example, to assess the drug effect in patients with immune

SECTION	ISSUE	PROPOSED CHANGE
	immune thrombocytopenic purpora, the disorder ascertained by operational definition International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis code 287.31 should be validated based on the conceptual definition of the disorder, which includes signs and symptoms, levels of platelets, and exclusion of other possible causes of thrombocytopenia. The ICD 10 codes are more commonly used in databases. Hence, suggest use of ICD-10-CM instead of ICD-9-CM) for this example. It is stated that "Key variables used to select the study population should be validated." However, unlike exposure, outcomes, and covariates, there is no discussion in the draft guidance around the Agency's expectations/considerations for the validation of inclusion/exclusion criteria. The draft draft guidance states that validation studies are needed. Information is missing here about how to link a certain indication to a drug exposure in order to establish one's cohort. In claims data, researchers cannot establish for certain if a particular drug was prescribed to treat any given indication. The researcher will have to document in detail in their study proposal how they will link their indication to their exposure drug. For example, a researcher wants to examine the association between fluoroquinolone antibiotics and the outcome of tendon rupture in a cohort of individuals prescribed their antibiotic for urinary tract infection. Researchers need	thrombocytopenic purpora, the disorder ascertained by operational definition International Classification of Diseases, Ninth Revision, Clinical Modification ~~(ICD-9-CM)~~ **(ICD-10-CM)** diagnosis code 287.31 should be validated based on the conceptual definition of the disorder, which includes signs and symptoms, levels of platelets, and exclusion of other possible causes of thrombocytopenia. BIO recommends that the final draft guidance provide citations of valid algorithms or use of datasets where data are consistently abstracted may be sufficient. BIO recommends that the expectations/considerations for the validation of inclusion/exclusion criteria be described and/or cross-referenced to the corresponding principles already covered for exposure, outcomes, or covariates that would apply to inclusion/exclusion criteria as well. BIO recommends that in this paragraph, FDA might consider adding 1-2 sentences emphasizing the need for researchers to describe in detail how they are going to not only define their cohort based on an indication of interest, but how they are going to link said indication to their exposure drug of interest in scenarios where a particular exposure is prescribed for said indication. BIO recommends that the Agency clarify the context and scope of the statement "Key variables used to select the study population should be validated".

SECTION	ISSUE	PROPOSED CHANGE
	to explain how they are identifying those fluoroquinolone users who were given their prescription to treat UTI. Perhaps if ICD-9 codes for UTI are present in the 14 days prior to their antibiotic prescription, this might indicate that the fluoroquinolone was prescribed to treat UTI. Typically, selection criteria in a comparative study if they apply similarly in both groups do not bias the comparison. They may however limit the generalizability of the inference. Thus, it is unclear whether validation of the key selection criteria is at all necessary and should take priority over showing that misclassification, if it exists, will not bias the findings or limit too much the generalizability	
Lines 540-543	The draft guidance states, "The protocol should include a detailed description of methods for determining how inclusion and exclusion criteria…" Outside of traditional clinical research protocols, EHRs/medical claims data use terminology that may not be identified by the agency.	BIO recommends that the Agency clarify that they would accept terminology not represented in the FDA Data Standards Catalog (e.g., ICD-10, RxNorm, etc.)
Line 543-544	The draft guidance states, "The protocol should address the completeness and accuracy of the information collected in the proposed data source to fulfill the inclusion and exclusion criteria. " Information on completeness of data needed for implementing I/E criteria might not be always available at the time of protocol finalization. In the case of EHR when unstructured data is used for evaluating I/E criteria, this might not be feasible. Suggest specifying it could be noted in the analysis plan.	BIO suggests the following edit: "The protocol **and/or analysis plan** should address the completeness and accuracy of the information collected in the proposed data source to fulfill the inclusion and exclusion criteria. "

SECTION	ISSUE	PROPOSED CHANGE
Entire section	Paper refers extensively to misclassification.. reference to sources of bias as used commonly used in the literature could aid understanding.. e.g. Table 1 in the paper *The use of external controls: To what extent can it currently be recommended? Burger U. et. Al Pharmaceutical Statistics. 20 21;1– 15.*	BIO recommends that the Agency refer to types of Bias in sections of the document e.g. if the RWD was intended as an external control to data from a clinical trial.. refer to Selection bias
C. *Exposure Ascertainment and Validation*		
1. Definition of Exposure		
Lines 569-571	The draft guidance states, "The product of interest is referred to as *the treatment*, and may be compared to no treatment, a placebo, standard of care, another treatment, or a combination of the above." The draft draft guidance states that a possible comparator is placebo. It is unclear how placebo as a treatment group would be assigned in an EHR or claims database analysis.	BIO suggests deleting "placebo" or clarify how this would be defined in a database study: The product of interest is referred to as *the treatment*, and may be compared to no treatment, ~~**a placebo**~~, standard of care, another treatment, or a combination of the above.
Line 581-582	The draft guidance states, "This will usually require an understanding of the pharmacological properties of the drug…" Half-life of the compound will impact the duration needed to evaluate the effect, the MoA is important to account for biological plausibility of the comparison between different compounds and the endpoints used to evaluate the effects.	BIO recommends the following edit: "This will usually require an understanding of the pharmacological properties **(e.g. half-life) and MoA** of the drug …"
2. Ascertainment of Exposure: Data Source		
Lines 593-595	The draft guidance states, "The protocol should describe the coding system used, the level of granularity represented (e.g., using RxNorm mapping to the National Drug Code [NDC] identifiers), and the specificity attained by the coding system."	BIO recommends that the final draft guidance should provide guidelines for acceptable reference standards to evaluate coding systems.

SECTION	ISSUE	PROPOSED CHANGE
	BIO notes that reference standards for determination of specificity of coding system are not identified.	
Lines 597-599	The draft guidance states, "When relying on coded data, the operational exposure definitions should be based on the coding system of the selected data source and reflect an understanding of the prescription, delivery, and reimbursement characteristics of the drug (if applicable) in that data source." Some coding systems may change the code of a particular procedure/drug over time or a specific code may be general and encapsulate many more specific procedures/drugs, with a shifting distribution over time.	BIO recommends the following edit: "When relying on coded data, the operational exposure definitions should be based on the coding system of the selected data source and reflect an understanding of the prescription, delivery, and reimbursement characteristics of the drug (if applicable) in that data source **over time**".
Lines 599-602	The draft guidance states, ""For example, in the United States, the operational definition should include the appropriate pharmacy codes (NDC or Healthcare Common Procedure Coding System) to capture the use of the drug in various settings." Sometimes the Healthcare Common Procedure Coding System (HCPCS) use non-J codes (e.g. Aflibercept (J0178, c9291, Q2046)). FDA may want to consider this application and modify the sentence in Lines 600 - 601 "(NDC or Healthcare Common Procedure Coding System)" to "(NDC or Healthcare Common Procedure Coding System or other applicable codes)".	BIO recommends the following edit: For example, in the United States, the operational definition should include the appropriate pharmacy codes (~~**NDC or Healthcare Common Procedure Coding System**~~ **NDC or Healthcare Common Procedure Coding System or other applicable codes**) to capture the use of the drug in various settings
Lines 619-622	The draft guidance states, "Uncaptured prescriptions might include low-cost generic drugs, drugs obtained through discount programs, samples provided by pharmaceutical companies and dispensed by health	BIO recommends that the Agency provide a reference for this statement.

SECTION	ISSUE	PROPOSED CHANGE
	care providers, and drugs sold via the internet or patient out-of-pocket purchases."	
3. Ascertainment of Exposure: Duration		
Lines 629-630	The draft guidance states, "The data source should capture the relevant exposure duration (anticipated use of a product over time)." The draft guidance indicates that the "relevant exposure duration" should be captured in the proposed data source, but this does not allow for "intention-to-treat" approaches, where the goal is to measure outcomes associated with treatment initiation.	BIO recommends the following edits: "The data source should capture the relevant exposure duration (anticipated use of a product over time)**, when relevant to the treatment and scientific question of interest**."
Lines 632-643	The draft guidance states, "FDA recommends describing the duration of exposure as well as the period during which the exposure is having its effect relative to the outcome of interest. Duration may refer to continuous exposure or cumulative exposure, depending on the study question. " These sources of bias will be very difficult to address.	BIO suggests that this topic could be raised in a workshop where regulators and data and analytics organizations work to address this topic.
Lines 645-649	The draft guidance states, "Because patients may not refill their prescriptions exactly on time or, alternatively , may refill their prescriptions early, gaps or stockpiling in therapy may exist and may be reflected in the data. FDA recommends describing and justifying in the protocol how researchers will measure use, address potential gaps in therapy in the data source, and handle refill stockpiling if there are early refills." In addition to what the FDA has mentioned here (patients may not refill prescriptions on time or refill early), FDA might also want to mention adherence to medication use. Patients may not be adherent to their medication, which could extend the amount of time that	BIO recommends that the Agency consider adding a sentence after "…if there are early refills," which recommends describing and justifying in the study protocol any action being taken by researchers to address medication non-adherence, if possible. BIO recommends the Agency provide the example of defining a maximum allowable gap between consecutive prescriptions to use to define continuous exposure vs. a new treatment episode.

SECTION	ISSUE	PROPOSED CHANGE
	they are using a drug past what the days supply of the drug indicates in claims. Adherence cannot be captured in claims, however, investigators in some situations could describe in their protocols any actions being taken to address adherence. For example, extra days might be added onto the days supply of an exposure drug in order to account for potential non-adherence or improper medication use.	
4. Ascertainment of Exposure: Dose		
Line 656	The draft guidance states, "Data about exposure should include information about dose" More precise information, especially for drugs with various dosing frequency that can have an effect on e.g. safety should be specified.	BIO recommends the following edit: Data about exposure should include information about ~~**dose**~~ **dosing regimen (dose level and frequency).**
Entire section	Single dose exposure, especially for drugs used chronically, can lead to misclassifications.	BIO recommends that the Agency include the concept of minimum effective dose since ascertainment by using single dose exposure, especially for drugs used chronically, can lead to misclassifications.
5. Validation of Exposure		
Lines 668-671	The draft draft guidance states, "Other than for medications administered in hospital settings or infusion settings, electronic health care data capture prescriptions of drugs and the dispensing of drugs to patients, but generally do not capture actual patient drug exposure because this depends on patients obtaining and using the prescribed therapy." When EHRs do not have an on-site pharmacy or are linked with pharmacy data, they typically do not capture drug dispensing data. In addition, prescriptions may not be filled if rejected during the preauthorization process or if patients choose not to.	BIO suggests the following edits: Other than for medications administered in hospital settings or infusion settings, ~~**electronic health care**~~ **claims** data capture prescriptions of drugs and the dispensing of drugs to patients **while electronic health records may not capture drug dispensing.** ~~but~~ **Claims and EHR data** generally do not capture actual patient drug exposure because this depends on patients obtaining and using the prescribed therapy.

SECTION	ISSUE	PROPOSED CHANGE
Line 673-675	The draft guidance states, "Validation ideally involves a comparison of the exposure classification in the proposed data source with a reference data source." Additional guidance is needed on what the FDA considers as an acceptable reference data type and source. In addition, it is a monumental task to validate all variables (some may be self-explanatory and others may have become standard practice). FDA should provide more clarity on which variable or design elements (e.g., exposure) should go through validation and in which scenarios.	BIO recommends that the Agency provide clarification on whether patient surveys can be conducted to validate exposure if appropriate reference data sources are not available.
Lines 680-686	The draft draft guidance states, "For prescribed medications used in outpatient settings, dispensing or billing data would tend to be more accurate than most EHRs in reflecting exposure to a drug by documenting that the prescriptions were filled…." The language regarding when EHR versus claims data would more accurately capture certain types of medications depending on the healthcare setting does not take into account that this assessment may depend on the specific context at hand (type of medication, disease, data source type, etc.).	BIO suggests the following edit: "**In some instances**, for prescribed medications used in outpatient settings, dispensing or billing data ~~**would tend to**~~ **may** be more accurate…" BIO recommends that the Agency clarify if they are requesting that RWE studies be conducted in multiple databases to ensure robust insights are derived. If so, the Agency should articulate this as a general consideration in B. Selection of Study Population.
Lines 692-694	The draft guidance states, "FDA recommends documenting the methods used to calculate and validate duration, dose, switching, and other characteristics of exposure. Validation and misclassification issues should be addressed in appropriate study documents."	BIO suggests revising to include the evaluation of exposure(s) is necessary, but new validation studies are not expected if these studies have already been previously performed and published.

SECTION	ISSUE	PROPOSED CHANGE
	It is unclear whether FDA expects validation studies for exposure(s) – even if prior validation studies have been carried out and performance measures reported.	
Entire section	The draft guidance states "Validation ideally involves a comparison of the exposure classification in the proposed data source with a reference data source ..." One aspect of drug exposure that is worth consideration is how drugs are actually taken by patients after they are filled in the pharmacy setting (i.e. adherence). BIO notes that deficiencies in defining the index date or the censoring criteria related to end of therapy were noted in recent reviews by the oncology divisions of applications containing EHR data. It would be useful for the agency to clarify their draft guidance relative to these important design elements.	BIO recommends that the Agency clarify that if no appropriate reference data are available if sponsors should used previous relevant publication or are there other options that the Agency recommends. BIO recommends that the Agency consider including in Subsection 3 discussions around ascertainment of the start and end of exposure therapy.
6. Dosing in Special Populations		
Lines 701-703	The draft guidance states, "For example, in assessing dosing in patients taking drugs with substantial renal clearance, it may be necessary to have access to measurements of serum creatinine, creatinine clearance, or estimated glomerular filtration rate to assess appropriateness of dosing." The example provided seems less relevant to a non-interventional study where the goal is to measure real-world treatment effect based on how the medication is being used in routine clinical practice.	BIO suggests the following edit: "~~For example, in assessing dosing in patients taking drugs with substantial renal clearance, it may be necessary to have access to measurements of serum creatinine, creatinine clearance, or estimated glomerular filtration rate to assess appropriateness of dosing~~."
7. Other Considerations		
Lines 721-725	The draft guidance states, "A study's definition of concomitant medication use should be described in detail. Definitions of concomitant medication use might	BIO recommends that the final draft guidance address exposure collection other than those for the investigational product of interest, such as combination therapy components,

SECTION	ISSUE	PROPOSED CHANGE
	include instances when drugs are dispensed on the same day, when drugs have overlapping days' supply, or when patients have filled prescriptions for two or more drugs during the study period. Limitations to ascertainment of concomitant drugs (e.g., nonprescription drugs) should also be described." In many clinical study settings, exposure to investigational product, to standard of care therapy, to combination therapy (e.g., chemotherapy backbone in add-on therapies), and rescue therapy may need to be collected in detail separately. The exposure to treatments other than investigational product of interest provides important context to understand and assess treatment effect.	standard care therapy for add-on therapies, rescue treatment, etc. BIO also recommends that the Agency provide clarification and examples of when concomitant medication use needs to be described as this may not be necessary for studies where causal effect of treatment is not the main study aim.
Line 714	The draft guidance states, "… the time period (if the comparator group is not concurrent with the treatment group)."	BIO recommends that the Agency provide additional draft guidance on the length of the time period that can be extrapolated back to, when no concurrent comparator group is available.
D. *Outcome Ascertainment and Validation*		
Entire section	This section focuses on discrete outcomes or acute events as outcomes and mentions Mortality as an outcome.	BIO recommends that for completeness, this section may need to add continuous variables as outcomes as well as other time-to-event variables as outcomes. For time-to-event outcomes, censoring should be addressed. BIO also recommends that the agency clarify how the validation may be conducted when using time-to-event outcomes from RWD.
Lines 763-767	The draft guidance states, " Since achievement of an objective response (tumor shrinkage), …" We may be able to measure clinical trial endpoints in a small fraction of patients (and a small fraction of a large database may correspond to a large sample size). More consideration should be given to how these data	BIO recommends that the Agency expand the current discussion to comment on how data on clinical trial endpoints such as RECIST 1.1 can be leveraged if these are measured in a fraction of patients in the real-world dataset. BIO also recommends that the Agency consider providing more clarity around the validation needed for surrogate RWD endpoints, i.e., tumor assessment measures.

SECTION	ISSUE	PROPOSED CHANGE
	can be leveraged and their fitness-for-purpose assessed.	
1. *Definition of Outcomes of Interest*		
2. *Ascertainment of Outcomes*		
Lines 781-783 **Lines 807-810**	The draft guidance states, "…the protocol should provide a detailed description and rationale for the methods and tools used to process the unstructured data and the validation of those methods." The draft guidance also states, "The protocol should include a detailed description of the operational definition, the coding system, the rationale and associated limitations of information selected to construct the operational definition (e.g., selection of primary or secondary diagnosis codes for which the order may not correspond to their medical importance),...."	In several cases, the draft Draft guidance mentions that certain source data and algorithms should be provided in the protocol. This is most relevant as it pertains to AI-based extraction tools for unstructured data. These algorithms are proprietary in nature and the expectations outlined in the draft guidance would necessitate data providers sharing IP with drug sponsors. An alternative solution would be to provide descriptions of the source data and algorithms and then to have more detailed data available for FDA's review upon request or upon inspection.
Line 784-788	The draft guidance states, "When patient- or physician-generated data (e.g., data required for subjective end points) are proposed to assess the outcome of interest…"	BIO recommends that the Agency consider referencing current COA draft guidance on validation of patient- or physician-generated data to assess outcomes.
3. *Validation of Outcomes*		
Line 820	The draft guidance states, "FDA expects validation of outcome variable to minimize outcome misclassification." The draft draft guidance describes that validation studies are needed.	BIO recommends that the final draft guidance provide citations of valid algorithms or use of datasets where data are consistently abstracted may be sufficient.
Lines 820-827	The draft guidance states, "Outcome validation involves using a clinically appropriate conceptual outcome definition to determine whether a patient's status,	BIO recommends that the Agency clarify whether there are more than two options for validation of outcomes.

SECTION	ISSUE	PROPOSED CHANGE
	classified by an operational definition, truly represents the outcome of interest, typically by reviewing clinical details recorded in the patient's medical records in either electronic or paper format." Two options are highlighted: complete verification of the outcome variable, and assessment of the operational definition in validation studies (which would presumably require access to a reference data source). If complete verification is infeasible and no suitable reference data source exists, it is unclear how to proceed.	BIO recommends the Agency consider EHRs where there is no original medical record (i.e., the individual heath record is the de-facto source data)
Lines 826-827	The draft draft guidance states, "… outcome of interest, typically by reviewing clinical details recorded in the patient's medical records in either electronic or paper format." Externally accepted gold standard data sources (NDI for example) should also be highlighted.	BIO recommends the following edits: " outcome of interest, typically by reviewing clinical details recorded in the patient's medical records in either electronic or paper format **or by using accepted external sources (e.g., NDI)**."
Line 835	The draft guidance states, "Reporting of comparison metrics (e.g., kappa statistic)…" Adding a reference to Cohen's kappa statistics would help the reader to investigate these measures.	BIO recommends that the Agency provide a reference to Cohen's kappa statistic.
Line 836-838	The draft guidance states, "An estimated medical record retrieval rate should be justified in the protocol, and the implications for internal and external validity should be discussed." It is unclear what the medical record retrieval rate would be used for or what is meant by it (i.e., is it about completeness of medical record at the individual level,	BIO recommends that the Agency clarify what the medical record retrieval rate is and what it would be used for.

SECTION	ISSUE	PROPOSED CHANGE
	or at study cohort level?).	
Lines 838-843	Blinding of abstractors Eliminating the potential for observer bias when using unstructured data by blinding abstractors is operationally difficult and has limited feasibility.	BIO requests that FDA consider alternative approaches to the use of blinding abstractors due to the operational difficulty and limited feasibility of the approach.
Lines 842-843	The Draft guidance states: "The protocol should provide a description of how observer bias will be handled". However, unlike other similar recommendations in the draft guidance, which are accompanied with some background narrative around the key considerations, there are no other text/sentences discussing observer bias.	BIO recommends that the Agency provide some background narrative to further illustrate the key considerations around observer bias related to this recommendation in the draft guidance.
Lines 845-846	The draft guidance states that "…complete verification of the outcome variable, each subject is assigned an accurate value of the outcome variable…" It is unclear how an accurate value is assigned to each subject and if this refers to a specific type of statistical or sensitivity analysis.	BIO recommends that the Agency provide clarity on how an accurate value is assigned to each subject and if this refers to a specific type of statistical or sensitivity analysis.
Line 847-848	The draft guidance states, "In practice, a more commonly used approach is to assess the performance of an operational definition in validation studies." It is unclear if it would be acceptable to use the same data for validation as is used to answer the study question.	BIO recommends that the Agency clarify whether these validation studies should be planned within the same study or in a separate standalone study.
Lines 848-851	The draft guidance states, "Performance measures, such as sensitivity, specificity, and predictive values, do	BIO recommends that the final draft guidance should provide acceptable ranges for performance measures.

SECTION	ISSUE	PROPOSED CHANGE
	not accurately classify cases and non-cases; rather, they inform the degree of outcome misclassification and facilitate the interpretation of results in the presence of misclassification." We note that standards for performance measures are not specified.	BIO recommends that the Agency consider mentioning the approach where predictive algorithms can be developed based on validation analyses and can be used to classify cases and non-cases. One example is: *Esposito DB, Banerjee G, Yin R, Russo L, Goldstein S, Patsner B, Lanes S. Development and Validation of an Algorithm to Identify Endometrial Adenocarcinoma in US Administrative Claims Data. J Cancer Epidemiol. 2019 Nov 3;2019*
Lines 855-859	The draft guidance states, "When the concern with false-negative cases is negligible (e.g., when the sensitivity is deemed sufficiently high so that the number of false-negative cases is minimal), a high PPV might be adequate to provide confidence in the validity of the outcome variable, whereas a moderate-to-low PPV might warrant complete verification of the outcome variable for all potential cases." It is unclear what FDA considers "high" PPV.	BIO recommends that the final draft guidance provide acceptable ranges for performance measures.
Lines 864-866	The draft guidance states, "Overall, the required extent of validation should be determined by necessary level of certainty and the implication of potential misclassification on study inference." It is unclear how the necessary level of certainty is determined for different contexts of use. An example could help sponsors understand FDA's thinking.	BIO recommends that the Agency clarify how the necessary level of certainty is determined for different contexts of use and provide an example to help sponsors understand the Agency's thinking.
Lines 875-879	The draft guidance states, "Because missing true cases is particularly a concern for infrequently reported outcomes, one approach is to select an operational definition of high sensitivity and perform complete verification of the outcome variable for all potential cases to maximize the likelihood that the true cases are	BIO recommends the following edit: "Because missing true cases is particularly a concern for infrequently reported outcomes, one approach is to select an operational definition of high sensitivity and perform ~~**complete**~~ verification of the outcome variable ~~**for all potential cases**~~ **to the level of precision needed to provide a reliable and clinically-meaningful estimate of treatment effects** to maximize the

SECTION	ISSUE	PROPOSED CHANGE
	all identified and that false-positive cases are minimized through validation." Statement of "all potential cases" seems beyond the stated intent of other sections (e.g., line 447 – "determining what variables of interest might require validation and to what extent"; line 823 – "assessing the performance of the operational definition of the outcome might suffice"). Large contents of the draft guidance are focused on validity, and yet recording each data point with as much precision as possible will not necessarily result in a more reliable estimate of treatment effects.	likelihood that the true cases are all identified and that false-positive cases are minimized through validation." BIO recommends that the Agency consider expanding verification to the concept of reliability.* * National Academies of Sciences, Engineering, and Medicine 2019. Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series. Washington, DC: The National Academies Press. https://doi.org/10.17226/25352 fROM PRECISION TO RELIABILITY pp. 52 – 55 Robert M. Califf
Lines 900-903	The draft guidance states, "Without complete patient information and complete verification of the outcome variable…" "Complete" patient information is often infeasible, particularly in the context of RWD.	BIO suggests the following edit: "Without ~~complete~~ **adequate** patient information and complete verification of the outcome variable…"
Lines 925-929	The draft guidance states, "For example, the physician who observed, diagnosed, and documented whether or not an outcome occurred could have been the same physician who made a decision as to which patients received the treatment meant to prevent that outcome, or the physician could have monitored disease progression or treatment side effects differently, given the knowledge as to which treatment they received." The example provided seems less relevant to EHR or claims database studies and more relevant to a prospective interventional study.	BIO suggests removing this example: ~~**"For example, the physician who observed, diagnosed, and documented whether or not an outcome occurred could have been the same physician who made a decision as to which patients received the treatment meant to prevent that outcome, or the physician could have monitored disease progression or treatment side effects differently, given the knowledge as to which treatment they received."**~~ BIO recommends that the Agency consider that the final draft guidance could highlight the importance of study investigators

SECTION	ISSUE	PROPOSED CHANGE
		being blinded to treatment assignment when designing a database study and making decisions throughout the analytic phase. See: McGrath et al. Lessons Learned Using Real-World Data to Emulate Randomized Trials: A Case Study of Treatment Effectiveness for Newly Diagnosed Immune thrombocytopenia. *Clin Pharmacol Ther* 2021. Online ahead of print.
Lines 943-946	The draft guidance states, "Regarding outcome validation, sponsors should justify the proposed validation approach, such as validating the outcome variable for all potential cases or non-cases, versus assessing the performance of the proposed operational definition; if the latter will be done, justify what performance measures will be assessed."	BIO recommends that the Agency include an Appendix with specific examples of validation approaches, thresholds of sensitivity/specificity, etc.
Entire section	It is currently unclear from the current text what a validation study would look like. In Lines 943 – 946, a distinction is drawn between "validating the outcome variable for all potential cases or non-cases versus assessing the performance of the proposed operational definition", although it is currently unclear what the key difference is between the processes of validation and performance assessment.	BIO recommends that the Agency clarify what the design of a study intended to validate the operational definition of an outcome might look like.
4. *Mortality as an Outcome*		
Lines 976-979	The draft guidance states, "These patients should be included in searches of vital statistics systems to see whether their absence (disenrollment) from the system is because of death, and it may be necessary to classify their deaths as an outcome of interest in the absence of data to the contrary." The CDC's NDI is the gold standard database for mortality and cause of death. Because these data can only be accessed through a formal application to the NCHS, this challenge should be acknowledged given	BIO recommends that the Agency acknowledge and provide draft guidance on the fact that it may be infeasible for sponsors to link EHR data to vital statistics given data privacy issues that commercial RWD data vendors need to adhere to. It may be possible for the database owner to conduct the search and provide the information to the Sponsor as supplemental information. BIO recommends that the Agency provide additional draft guidance on whether censoring or assigning death to patients who are lost to follow up depends on the circumstances of the

SECTION	ISSUE	PROPOSED CHANGE
	the time needed to conduct such a validation study as well as the lack of recent data (e.g., 18-month data lag time). There are also concerns with patient identification when linked to vital indexes such as NDI. Assuming patients who are lost to follow up in EHRs are deceased in the absence of data to suggest otherwise may also lead to misclassification of the mortality outcome. Classifying patients who are lost to follow up in an EHR as deceased in the absence of data to the contrary, will result in a either a true positive or false positive death. In the case of a false positive death, the effect would be to bias overall survival rates downward and in RWD cohorts used as external comparators in a single arm trial or in hybrid trials, this would potentially exaggerate treatment effect in the trial experimental treatment arm. Patients who are lost to follow up in the EHR/Claims records following exposure may actually have died. It has been suggested to search these patients in vital records systems to ascertain death. This may be challenging to do without the necessary identifiable patient information (e.g., social security number) that are typically not accessible to researchers using secondary data for studies.	study (i.e., an RWD external comparator or hybrid study using RWD versus other types of study designs).
Lines 971-973	The draft guidance states, "If the death is not captured in the electronic health care data systems, patients who die after having been exposed to the study drug might be observed in electronic health care data."	BIO recommends that the Agency consider including some background on Estimands, and how intercurrent events may be handled in this section.

SECTION	ISSUE	PROPOSED CHANGE
Entire section	Other time-to-event outcomes and censoring should be addressed.	BIO recommends that the Agency consider including discussions on other time-to-event outcomes and censoring in this section.
E. *Covariate Ascertainment and Validation*		
Lines 983-984	FDA does not mention variables associated with exposure only, or instruments. These variables are important to consider and identify in a study protocol, especially when deciding what variables to control for or what variables to include in a propensity score. This might be discussed in the next FDA RWE draft guidance focused on study design and analysis, but if not some information could be added here. Confounders vs effect modifiers – some covariates can be both confounders and/or effect modifiers.	BIO recommends that the Agency consider mentioning instruments along with confounders and effect modifiers and the importance of identifying instrumental variables. Inclusion of a direct acyclic graph (DAG) in study protocols might be useful for researchers as they decide which variables to control for.
Entire section	Examples of confounders are given. However, this is not clearly distinguished from e.g., covariates. This section is focused on individual variable level ascertainment and validation. There is no draft guidance yet 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.	BIO recommends that the Agency consider defining confounders in Section E1 as well, and to add these definitions to the glossary.
1. *Confounders*		
Lines 1001-1002	The draft guidance states, "*FDA recommends considering potential linkages with other data sources or additional data collection to expand the capture of important confounders that are unmeasured or imperfectly measured in the original data source*." Other than potential linkage or additional data collection, it is	BIO recommends the following edits: "*FDA recommends considering potential linkages with other data sources or additional data collection to expand the capture of important confounders that are unmeasured or imperfectly measured in the original data source*. **If linkages with other data sources are not possible, the sponsor should discuss with the**

SECTION	ISSUE	PROPOSED CHANGE
	unclear if a sponsor could use quantitative analysis methods to provide evidence on the robustness of study results given the possible existence of unmeasured or imperfectly measured confounders. Linkage and additional data collection may not be always feasible. The draft 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.	**specific review division the appropriateness of using proxy variables to the specific unmeasured confounders in their study**."" BIO recommends that the Agency expands upon the draft guidance to clearly state whether quantitative analysis methods are an acceptable approach in which to address the possible existence of unmeasured or imperfectly measured confounders. BIO recommends 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.
Entire section	The description of exposure should also comprise the description of the comparator time period in case of a comparison with "no treatment" or with placebo.	BIO recommends that the Agency expand the definition to periods with no treatment, treatment interruptions, previous treatments etc.
2. *Effect Modifiers*		
Lines 1011-1014	The draft guidance states, "The potential for effect modification by demographic variables (e.g., age, gender, race, ethnicity) or pertinent comorbidities should be examined in the study, and relevant effect modifiers should be available in the chosen data source." Biomarkers and genetic mutations are important effect modifiers and can be added to the Draft guidance. While effect modifiers ideally should be available in the chosen data source, it should be recognized that often all effect modifiers are not fully understood or fully available.	BIO suggests the following edit: "The potential for effect modification by demographic variables (e.g., age, gender, race, ethnicity) or pertinent comorbidities should be examined in the study, and relevant effect modifiers should **ideally** be available in the chosen data source**, and mitigation strategies presented where this is not possible**." BIO also recommends that the Agency include biomarkers and genetic mutations as important effect modifiers in the draft guidance.

SECTION	ISSUE	PROPOSED CHANGE
Entire section	There is a lack of draft guidance on assessing effect modifiers in both statistical and biological scales.	BIO recommends that the Agency include some information on assessing effect modifiers in both statistical and biological scales.
3. *Validation of Confounders and Effect Modifiers*		
Entire Section	BIO notes that it seems infeasible to validate every covariate	BIO recommends that FDA provide draft guidance for why/when covariates necessitate validation.
Lines1018-1019	The draft guidance states, "For all key covariates, including confounders and effect modifiers, FDA recommends providing and justifying the validity of operational definitions in the protocol and study report." Validation of all operationalized covariates may not be the most efficient means of addressing potential data quality issues.	BIO suggests including language that recognizes that important covariates which are also more prone to errors receive the highest priority with respect to validations.
Lines 1026-1034	The draft guidance states, "When evaluating the validity of covariate…" It is not clear how to proceed if no gold standard is available for validation. For example, would additional efforts be required to prospectively generate that data, such as conducting patient surveys?	BIO recommends that the Agency clarify what approaches are possible to validate covariate operational definitions when no suitable reference exists.
Lines 1036-1038	The draft guidance states, "When supplemental information is needed to capture important covariates or is used for covariate validation, FDA recommends describing the likelihood of obtaining the supplemental information for the overall study population." It appears the draft draft guidance is starting with a default position that supplementary information should be available for the entire population. In contrast, a prior FDA-sponsored study* supports that supplementary information for a small portion of study	BIO recommends the following edits: "When supplemental information is needed to capture important covariates or is used for covariate validation, FDA recommends describing the likelihood of obtaining the supplemental information ~~**for the overall study population**~~ **needed to provide a reliable of treatment effects**."

SECTION	ISSUE	PROPOSED CHANGE
	population can provide the necessary scientific basis for interpreting the study. *See Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using Real-World Data to Predict Findings of an Ongoing Phase IV Cardiovascular Outcome Trial: Cardiovascular Safety of Linagliptin Versus Glimepiride. Diabetes Care. 2019 Dec;42(12):2204-2210	
VI. DATA QUALITY DURING DATA ACCRUAL, CURATION, AND TRANSFORMATION INTO THE FINAL STUDY-SPECIFIC DATASET		
Entire section	BIO agrees with the importance of data quality and note that industry would benefit from more specific draft guidance in this area. Data vendors of claims and EHR do not typically share the level of documentation discussed in the draft guidance with sponsors. Sponsors do not always have access to subject-level data. It would be helpful for the agency to clarify to sponsors the level of expectation around documentation and data submission.	More specific draft guidance would be helpful so that sponsors and data providers can provide consistent data quality assessments that satisfy FDA's requirements. BIO believes that it would be useful for industry, data partners, and FDA to collaborate on developing a framework for data standards. If not completed prior to issuing the final draft guidance, it could be added later as an appendix to the final draft guidance. BIO recommends that the Agency clarify what the expectations relative to subject-level data submission of RWE are. The focus of the draft guidance is on documentation of the data quality and validation of the records. Specifically, it would be helpful to understand the Agency's thinking on if that takes priority over access and submission of subject-level data.
Lines 1073-1075	The draft guidance states, "The study protocol and analysis plan should specify the data provenance…" This draft guidance recommends that the data provenance be included in the protocol and analysis plan, however, in cases where the sponsor is using a vendor, some of those steps may be proprietary.	BIO recommends that the Agency clarify if separate documents from the sponsor and vendor to capture such information would be an option.
A. *Characterizing Data*		

SECTION	ISSUE	PROPOSED CHANGE
Lines 1086-1088	The draft guidance states, "The FDA recommends automated data quality reports that include the following characteristics and processes in a standardized way, when applicable to the chosen data source:"	BIO recommends that the Agency clarify if they recommend Sponsors submitting the data quality report(s) to the FDA as part of the protocol and statistical analysis plan review meeting.
Lines 1117-1120	The draft guidance states, "7. Any updates or changes in coding practices and versioning (e.g., International Classification of Diseases [ICD] diagnosis codes, Healthcare Common Procedure Coding System codes) across the study period that are relevant to variables of interest." Time-varying trends due to coding changes are mentioned a few times in the document.	BIO recommends that it would be helpful for FDA to address the Agency's interpretations of the coding trend (ICD-9 to ICD-10) analyses that are publicly available in Sentinel.
Line 1090	This section discusses data accrual. Data accrual is typically conducted by data vendors. Parts of the data accrual process described here, may be considered proprietary and therefore not possible to obtain from vendors.	BIO recommends that the Agency clarify its thinking on the role of data vendors in the data accrual process.
Line 1097 - 1099	Line 1097-1099 "*Provenance of core data elements to allow tracking of these elements back to their 1097 respective points of origin…*" For lots of EHR/claim databases, the data are de-identified before data submitted to a central data warehouse to create structured data (e.g., IBM MarketScan). If this is the case, it is very difficult, if not entirely impossible, to track the data elements back to their origins. Provenance of variables, and provenance of individual data values may require extensive and potentially proprietary information from vendors. More explicit draft	BIO recommends that the Agency provide more explicit draft guidance on what would constitute adequate provenance documentation.

SECTION	ISSUE	PROPOSED CHANGE
	guidance on what would constitute adequate provenance documentation is needed.	
Line 1129	Line 1129, "*Quality assurance (QA) testing and data quality checks employed across sites….*" It is unclear if this means that the study sponsor or data provider should establish a QA/QC standard and apply the standard to all the sites or if the sites could use their own QA/QC process.	BIO recommends that FDA provide clarification as to intent of the QA testing and data quality checks and the use of QA/QC standards such that the obligations of sponsors and data providers are clear.
Lines 1095, 1097, 1133	"Core data elements" has not been defined anywhere in the document. There's a reference of "key data element" in line 68. It is unclear if 'core data element' is the same as the 'key data element'.	BIO recommends that the Agency clarify the definition of core data elements.
Line 1145	The draft guidance states, "Conformance to open, consensus-based data curation standards, when applicable." Data vendors determine curation standards. Sponsors likely will not have much influence over curation standards.	BIO recommends that the Agency clarify their expectation of the sponsors role in determining curation standards.
Lines 1154-1198	As part of a multi-step data transformation process, traceability may be lost in this intermediary activity (refer to Figure 1). For example, if there are mappings for terminology or semantic harmonization (e.g., mapping between one standard to another) traceability may be lost.	BIO recommends that the Agency clarify if they want the intermediate processing of the data to be submitted with regards to quality and traceability. If so, BIO recommends that the Agency clarify what the expectation is on the level of detail and delivery format (e.g., report, log, etc.).
Lines 1182-1185	The guidance states that "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."	BIO recommends that the Agency provide examples or best practices for data linkage. For example, parameters indicating good quality of record linkage, and what patient information should be used for linkage in compliance with data privacy.

SECTION	ISSUE	PROPOSED CHANGE
B. *Documentation of the QA/AC Plan*		
Entire section	It is recommended that analyses be validated by double-programming, which is independent code development by at least two programmers, followed by comparison of analysis outputs, and investigation of all discrepancies. Code review alone, without double programming is generally not recommended. If double programming is not performed, the suggestion is to state as such in the Statistical Analysis Plan, with a description of alternative methods of analysis validation.	BIO recommends that the Agency discuss considerations around code review alone vs. double programming.
C. *Documentation of Data Management Process*		
Lines 1245-1248	The draft guidance states, "To facilitate FDA review, all submitted programs (e.g., those written by analysts) should be thoroughly annotated with comments that describe the intent or purpose of each data management and analysis step written in the program (e.g., annotate each data step in a statistical analysis program)." We recommend that the Agency clarify the extent of annotation needed and in what circumstances this will be required for submissions. It would be helpful to know if these are also required for post-approval safety studies, as they are not typically provided.	BIO suggests the following edit to clarify: "To facilitate FDA review**, analysis datasets, data definition tables, and programming codes (e.g., those written by analysts) for data derivation and data analysis related to the key objectives of the study (e.g., primary and key secondary endpoints) should be** ~~**all**~~ submitted **and** thoroughly annotated with comments that describe the intent or purpose of each data management and analysis step written in the program (e.g., annotate each data step in statistical analysis **programming codes**). " BIO recommends that the Agency clarify if the submitted programs are provided for FDA review if they are specifically requested by FDA or is this a general suggestion that all analysis programs written by analysts are to be included in the SAP. BIO also recommends the Agency clarify if this applies to all submissions, including submissions for post-approval safety studies. BIO recommends that the Agency clarify if there is any specific data structure recommended (like ADaM data is required for RCT).

SECTION	ISSUE	PROPOSED CHANGE
		BIO also recommends that the Agency clarify the documentation that should be provided for analyses performed within platforms (e.g., Aetion and other data platforms).
Lines 1234-1235	The draft guidance states, "All manual and automated data retrieval and transformation processes should be thoroughly assessed from data collection through writing of the final study report to ensure data integrity."	BIO recommends that the draft guidance indicate that in the case of 3rd party vendors when the sponsor does not have access to the source data, the data owner would have a responsibility to provide the details of the transformation process.
VII. GLOSSARY		
Line 1280 and Lines 1323	Line 1280, "confounder": The definition of "confounder" is inconsistent in the scientific communities and there is no consensus up to date. However, there is a publication that discussed and evaluated all those definitions and proposed a reasonable definition of "confounder". VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013;41(1):196-220. doi:10.1214/12-aos1058 The current definitions provided are not self-explanatory.	BIO recommends that the Agency consider the alternative definitions of confounder and refine the current definitions of confounders/effect modifiers in the glossary..
Line 1343	The current definition of missing data given in the Glossary is inconsistent with the definition given in ICH E9 Addendum. Alignment would help the reader.	BIO recommends that FDA align the definition of missing data with that given in the ICH E9(R1) Addendum
VIII. REFERENCES		