Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document contact (CDER) Scott Goldie at 301-796-2055 or (CBER) Office of Communication, Outreach, and Development, 800-835-4709 or 240-402-8010.

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

November 2018
Drug Safety

Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products Guidance for Industry

Additional copies are available from:

Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353
Email: druginfo@fda.hhs.gov
https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

and/or

Office of Communication, Outreach and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave., Bldg. 71, Room 3128
Silver Spring, MD 20993-0002
Phone: 800-835-4709 or 240-402-8010
Email: ocod@fda.hhs.gov
https://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

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

November 2018
Drug Safety

TABLE OF CONTENTS

Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products Guidance for Industry[1]

I. INTRODUCTION

This document provides guidance to applicants submitting investigational new drug applications (INDs), new drug applications (NDAs), biologics license applications (BLAs), or supplemental applications on the use of meta-analyses of randomized controlled clinical trials (RCTs) to evaluate the safety of human drugs or biological products within the framework of regulatory decision-making.[2] This guidance is also intended for FDA reviewers and for third-party entities that prepare or evaluate meta-analyses assessing the safety of drug products. Specifically, this guidance describes the factors FDA intends to consider when evaluating the strength of evidence provided by a meta-analysis studying the safety of drugs.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. BACKGROUND AND SCOPE

Evaluating the safety of drug products, both before approval and after marketing, is a fundamental responsibility of the FDA. This evaluation often requires combining and summarizing information from multiple sources, and meta-analysis is a useful tool for this purpose. The term *meta-analysis*, as used in this document, refers to the combining of evidence from relevant studies using appropriate statistical methods to allow inference to be made to the population of interest. The most common reason for performing a meta-analysis is to provide an estimate of a treatment effect or measure of relative risk associated with an intervention and to quantify the uncertainty about the estimated effect or risk, when data from a single existing study are insufficient for this purpose, and the conduct of a new, large study would be impractical, take

[1] This guidance has been prepared by the Office of Biostatistics in the Center for Drug Evaluation and Research, in cooperation with the Center for Biologics Evaluation and Research, at the Food and Drug Administration.

[2] For the purposes of this guidance, all references to *drugs* include both human drugs and biological products unless otherwise specified.

too long, or be unethical. The term meta-analysis sometimes refers to the quantitative synthesis in a systematic review (Cochrane Handbook 2011) and the term systematic review refers to the broader effort, including defining objectives and selecting and evaluating studies, as well as synthesis. We use the term meta-analysis more broadly to include consideration of study selection as well as overall design issues such as prespecification and reporting.

Unless a randomized controlled clinical trial is prospectively designed with a particular safety outcome as its primary endpoint and sized accordingly, the trial may not have sufficient sample size to detect important adverse consequences of drugs and to reliably evaluate whether there is increased risk of such events. This is because most serious drug-induced adverse events (1) are rare or (2) occur at only slightly increased frequency compared to background rates and are not obviously drug-related (e.g., cardiovascular events, cancers). Meta-analysis is most useful in the latter case, to detect and quantify an increased risk over the background rate of the safety event. For the former case, when events are rare and not expected to occur in the target population, meta-analyses may still be useful for improving the precision of the estimate of risk.

Meta-analysis factors into FDA's evaluation of potential safety issues in a variety of ways:

- Meta-analyses may be conducted by sponsors and submitted to FDA as part of an IND, NDA, BLA or supplemental submission.
- FDA may ask a sponsor to conduct a prospective meta-analysis, as it has recommended for sponsors of new antidiabetic therapies to treat type 2 diabetes in the draft guidance for industry, *Diabetes Mellitus – Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes*.[3]
- FDA may initiate its own meta-analysis in response to safety signals that FDA is aware of, using study data FDA has access to, but that may be unavailable to sponsors and other researchers. These meta-analyses typically have prospective protocols to address issues of bias and multiplicity, as discussed later in this document.
- FDA may evaluate a meta-analysis conducted by an external party that raises a safety concern about a marketed product.

Because regulatory actions may stem from a meta-analysis, it is important that rigorous principles be applied to such studies. In this guidance, the important principles underlying best practices for safety meta-analysis and the way that FDA intends to factor adherence to those principles into its decision-making are described. An overview of the most important principles presented in this guidance is as follows:

- Prespecification and transparency are recommended, as they enable a thorough evaluation of the meta-analysis.
- The criteria for selecting which trials to include should be determined prior to conducting the meta-analysis. The selection of the studies should not be based on the

[3] We update guidances periodically. To make sure you have the most recent version of a guidance, check the FDA Drugs or Biologics guidance web pages at: https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. https://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

trial outcomes, but rather on trial quality and consistency of critical design elements, and should be executed by parties masked to the outcomes of the trials, whenever possible.

- The quality and relevance of the individual trials and the quality of the trial data are critical determinants of the quality of the meta-analysis itself. Outcome ascertainment and adequacy of exposure periods are two of the most important determinants of trial quality.
- Meta-analysis conducted to meet safety objectives often requires re-purposing trials that were originally designed to meet efficacy objectives. This can be challenging, particularly if subject-level data are not available.
- Meta-analysis based solely on published trials is particularly problematic because of the potential for bias and error, both known and unrecognized.
- Generally accepted principles of good statistical practice should be followed in selecting the statistical methods to be used for meta-analysis (but this guidance is not prescriptive as to the choice of method).

This guidance applies to meta-analyses conducted in both pre-market and post-market settings. In the pre-market setting, the number and scope of trials may be limited, because the drugs are not yet approved for marketing, and these limitations may affect the ability to address the safety question of interest. In the post-market setting, the number and variety of trials available for inclusion are usually larger, as is the number of parties able to conduct the meta-analysis. In both pre- and post-market settings, the important principles guiding a well-planned and well-executed meta-analysis apply.

This document focuses specifically on meta-analyses conducted for purposes of safety evaluation using data from RCTs. Meta-analyses conducted to evaluate a product's effectiveness, either overall or within specific subgroups, are occasionally of interest to FDA, but the primary use of meta-analysis in the regulatory setting is for assessment of product risk. While meta-analyses of non-randomized studies may be informative for the assessment of certain safety outcomes, the issues related to such a meta-analysis are more complex, and the interpretation of the results more controversial. Meta-analyses of observational studies are therefore not addressed in this guidance.

Meta-analyses are conducted for both exploratory and confirmatory purposes. The primary focus of this guidance, however, is on meta-analyses with predefined hypotheses that are designed to confirm a suspected risk associated with a drug rather than on exploratory meta-analyses.

The subsequent sections of this guidance provide a detailed discussion of the important elements used in evaluating meta-analyses for regulatory purposes. In section III, the importance of the quality and relevance of the component trials included in a meta-analysis and the quality of the data from those trials are discussed. In section IV, the importance of prespecification and transparency in designing, conducting, and reporting a meta-analysis is described. In section V, the use of recommended statistical methods is discussed. In section VI, we summarize these technical considerations and discuss how they may be factored into a regulatory decision.

Section VII provides two examples illustrating the range of meta-analyses conducted for safety evaluation and FDA's use of the evidence provided by each.

III. THE QUALITY AND RELEVANCE OF CANDIDATE TRIALS

A. Basic Principles

Deciding what trials to include in a meta-analysis is an important step in the design and conduct of a high-quality meta-analysis. The major determinants for this decision should be the quality and relevance of the individual trials and the data collected in those trials. The component trials of a meta-analysis should be able to address the safety objectives of the analysis and be of sufficient quality to provide evidence useful for regulatory decision-making. The following are important factors to consider in determining whether the individual trials and associated data are of sufficient quality and relevance to ensure the validity of the meta-analysis:

- The extent to which the component trials are consistent with established standards for the design and conduct of adequate and well-controlled clinical trials
- The quality and completeness of safety outcome ascertainment in each trial
- The appropriateness of exposure and follow-up periods for estimating risk
- The appropriateness of the component trials' inclusion/exclusion criteria for defining the population at risk
- The appropriateness of the comparator used in each trial and of the doses for the test drug and comparator
- The relevance of the candidate trials to current medical practice
- The availability of subject level data from each trial

These factors are discussed further in the subsections that follow.

B. Consistency with Standards for Adequate and Well-Controlled Trials

The knowledge base, literature, and published guidelines for designing, conducting, and analyzing well-controlled clinical trials to demonstrate efficacy in support of an NDA or BLA are extensive and well-known (see, e.g., *E9 Statistical Principles for Clinical Trials*, International Council on Harmonisation (ICH) of Technical Requirements for Pharmaceuticals for Human Use). The same principles apply to the individual component trials of a meta-analysis, and the extent to which the component trials satisfy these principles has strong bearing on the quality of the meta-analysis to which they contribute. Notably, however, trials that are well-designed to measure the effect of a drug on a particular efficacy outcome may not necessarily be well-designed to measure an effect on another outcome, particularly an uncommonly occurring safety outcome, as discussed further in section III.C.

Some study designs may cause a candidate trial to be discouraged from inclusion in the meta-analysis. For example, randomized withdrawal studies, in which all subjects initially receive the drug and are then randomized to either remain on the drug or withdraw to a placebo or active control drug, may not be recommended for a safety meta-analysis. In these studies, subjects who cannot tolerate the test drug are excluded from the randomized portion of the trial, and the study

population may therefore not accurately represent the population at risk. Additionally, depending on the period of exposure needed for the adverse effect to occur, the initial exposure to the drug may result in events in both randomized groups and an underestimate of the relative risk. Crossover studies in which subjects receive different treatments at different periods of time may not be recommended for evaluating safety outcomes, if exposure to a treatment in one period can result in an adverse event occurring in a later period. Washout periods for safety outcomes may need to be longer than for efficacy outcomes. Other non-standard study designs such as enrichment trials, trials with add-on therapies, adaptive trials, and trials stopped at interim may raise similar issues.

C. Outcome Definition and Ascertainment

A high-quality meta-analysis has a carefully defined outcome variable with appropriate ascertainment procedures prospectively implemented in the component trials such as specific protocol-defined procedures for data collection and adjudication of safety outcomes. For example, if the outcome of interest is myocardial infarction, the protocol might instruct the investigators to collect laboratory and electrocardiogram data for suspected events during the trial. The results of these procedures might then be subject to adjudication by an independent panel to strengthen the evidence that a case event is real. Such procedures, however, are used primarily to assess effectiveness outcomes (does the treatment reduce myocardial infarction rates) and are not commonly used to assess safety outcomes, unless there is a specific concern known and planned for prior to study start (e.g., cardiovascular outcomes in studies of Type 2 diabetes drugs; suicidal events in studies of antidepressant drugs). Although prospective collection and adjudication of safety outcomes are desirable, they are usually not feasible, particularly in the most common setting of evaluating a new, unanticipated safety signal with data from trials already completed.

When the component trials are not prospectively designed to produce accurate ascertainment of the meta-analysis safety outcome, retrospective identification and adjudication of events will usually be recommended. In this situation, the safety outcome of interest should be clearly defined, and the identification and adjudication of events should be performed while masked. For example, in the antidepressants and suicidal events meta-analysis (section VII, Example 1), where suicidality was not specifically assessed, predefined search criteria were applied to adverse event data collected in the component trials. Based on the results of the search, narratives of candidate events were created, and a group of experts masked to treatment assignment classified the events into validated suicidal outcome categories. This resource intensive effort required subject-level data not directly available in the original trial datasets. A detailed meta-analysis protocol was developed that described the procedures necessary for obtaining and adjudicating the outcome data of interest prior to implementing those procedures.

Measurement bias (such as an over- or under-estimation of the rate of events because of imprecise or individualized interpretation of adverse event reporting) factors into determining whether outcome ascertainment is sufficient for a high-quality meta-analysis. Biases common to both treatment and control groups can occur when an outcome variable does not accurately represent the safety outcome of interest (e.g., is not specific enough, causing many irrelevant events to be reported, or is too narrowly defined, causing many events to be missed). Both

reporting problems may result in reduced power or a biased effect measure, but will not completely eliminate the ability to detect an effect. Of greater concern are reporting problems that can affect treatment groups differently, as they can eliminate the ability to detect an effect when one exists or that create the appearance of an effect when one does not exist.

Biased ascertainment of outcomes is one important concern in unmasked trials, where investigators or subjects may unconsciously, or consciously, under- or over-report medical events based on the known treatment assignment. Even in double-masked trials, there is a potential for differential bias to occur in safety reporting, especially when safety outcomes were not of primary interest in designing the trial. For example, a drug may cause discoloration of the urine, which in turn may lead to more evaluations and subsequent diagnoses of kidney disease. If anticipated, the trial protocols could have included an evaluation for kidney disease at scheduled times during the trials, thereby reducing the potential for biased reporting of that safety outcome.

Several strategies should be considered to minimize the impact of measurement bias. The use of safety outcomes that can be diagnosed readily and unambiguously, often called *hard outcomes,* can help minimize bias due to outcome ascertainment in a meta-analysis. For example, if vital status at the end of the study is known for all patients in all of the component trials, then use of death as the safety outcome effectively avoids any potential for ascertainment bias. If ischemic cardiovascular outcomes are of interest, ascertainment of myocardial infarction and stroke will be less prone to ascertainment bias than less specific events such as transient ischemic attack or angina. Excluding the less specific events or events that are difficult to ascertain objectively will probably reduce the power of the meta-analysis to detect a safety signal as well as the precision of the risk estimate that results, but the reduction in ascertainment bias may outweigh these losses. Precision and power can be quantified and reported with the meta-analysis results, whereas bias is typically unknown and difficult to measure. In general, reducing bias in a meta-analysis should be given greater weight than increasing precision and power.

It is important to define the period within which the safety outcome of interest is to be measured. For example, a safety outcome corresponding to the occurrence of anaphylactic events may call for the primary focus to be placed on the period of initial drug exposure, with a secondary focus on the entire drug exposure period. Including events beyond the initial exposure period may result in underestimation of the risk attributable to the drug. In cases where it is known that the effect of the drug diminishes when the drug is stopped, it might be recommended for the primary analysis to count outcomes only during the time a subject is on the drug (such as an on-treatment analysis).

Ideally, outcome definition and ascertainment should be as uniform as possible across the component trials. Trial-to-trial differences can introduce heterogeneity in safety outcomes, increasing the variability of the meta-analytic estimate of risk. Differences in outcome definition and ascertainment may be confounded with other trial design or subject population characteristics, making observed differences in risk measures difficult to interpret.

Outcome definition and ascertainment are particular problems for meta-analyses that rely exclusively on published trial data. Information taken from published articles about the component trials may be incomplete or lack specificity. Publications may not report on the safety

outcome of interest, and even when the outcome is reported, important details may be lacking, including whether the event occurred on or off randomized treatment and whether the outcome was defined a priori and uniformly across trials. Protocols, study reports, and subject-level data from the component trials are often important to determine whether the trial outcomes are adequate for supporting a high quality meta-analysis.

The definition of the safety outcome, the source data and any adjudication procedures that may have been employed should be prespecified in the meta-analysis protocol and consistently applied to all component trials, if possible (see Section IV.B).

D. Duration of Exposure and Length of Follow-Up

The duration of exposure and length of follow-up for each of the candidate trials should be factored into the criteria for trial inclusion. For an outcome with delayed appearance, such as cancer or bone injury, the inclusion of short-term trials may not be recommended. When subject-level data are available, analysis methods can be used to identify and account for differences in trial duration across studies (see Section V). Without subject-level data, it may not be possible to account for differences in duration, depending on the level of detail provided by the summary information available from each trial, and some trials may need to be excluded as a result.

Subjects prematurely stopping assigned drug or withdrawing from the trial can affect the comparability of subject groups with respect to safety outcomes ascertained over the course of the treatment or study period. The dropout pattern may result in dissimilar observation time between the two groups, resulting in more opportunity to observe the safety outcome in one group compared to the other. Simple adjustments for person-time of observation may not be sufficient to correct for non-comparability, because these adjustments assume constant hazards across time. The risk of the event may not be constant over time if, for example, the safety outcome tends to occur either early or late during treatment. Time-to-event analysis may also be insufficient if the dropout rates are indicative of informative censoring; for example, if the adverse events resulting in early discontinuations are similar to or predecessor events of the safety outcome.

When reviewing the component trials of a meta-analysis, it is important to consider the possibility of differential follow-up and informative censoring. Examining summary statistics and graphics by subject group of on-assigned drug time and follow-up time is usually helpful for this purpose, as is an examination of the stated reasons for stopping assigned-drug or discontinuing participation in the trial by subject group. The criteria for excluding individual trials for these reasons should be specified a priori and described in the meta-analysis protocol and analysis plan (see Section IV). If incorporated in the trial inclusion and exclusion criteria (applied to determine the component trials of the meta-analysis), a review to identify differential dropout rates should be performed masked to the safety outcome measurements. Regardless of the decision on inclusion, data summaries should be provided in the meta-analysis report to permit consideration of these issues.

E. Subject Populations

Wherever possible, the subject population for component trials should reflect the patient population hypothesized to be adversely affected by the drug. For cardiovascular safety outcomes, for example, trials that enrolled subjects with pre-existing cardiovascular risk factors may improve the ability of the meta-analysis to detect any risk associated with the drug. Conversely, including trials that excluded subjects with certain risk factors may limit the ability to detect risk. The inclusion/exclusion criteria of the component trials should be reviewed to determine if the corresponding subject populations are consistent with the objectives of the meta-analysis.

F. Dosing and Comparators

Although uniformity of dosing regimens and therapeutic indications studied across component trials is desirable, it may be that trials including other doses or conducted in other indications can contribute to the meta-analysis. For example, in some circumstances, it may be assumed that if a safety event is not observed at doses higher than the dose or doses approved, it should not occur at the approved dose or doses. In this scenario, including trials with dosing higher than the approved dose might be used to rule out an association. Information on dose response relationships may also support a possible relationship between drug use and safety outcomes. Similarly, including trials for indications outside the indication of specific interest may be useful in a safety meta-analysis, if it can be assumed that the association would not depend on the indicated use. Such assumptions can be examined to some extent through sensitivity analyses conducted on subsets of trials at particular doses or in particular indications (see Section V.D).

The suitability of the comparator drugs in the candidate trials should also be factored into the meta-analysis inclusion criteria. In some situations, the ideal comparator is a placebo, since a placebo cannot cause the safety outcome under investigation. However, placebo-controlled trials may not be feasible or ethical in certain disease areas. If trials with active drug comparators are used, attempts should be made a priori to determine if the active comparator is associated with the safety outcome of interest. The protocol specifications for concomitant therapy in the individual trials are also relevant, since concomitant therapies may be associated with the safety outcome.

G. Relevance to Current Medical Practice

Changes over time in the practice of medicine may affect the usefulness of some trials for contributing data to a meta-analysis. Older trials may no longer be relevant, if medical practice has changed such that current practices are able to prevent or reduce the occurrence of the safety outcome under investigation. Sensitivity analyses can be used to examine estimated risks as a function of the dates the component trials were conducted to determine if calendar trends pose a problem.

H. Availability of Subject-Level Data

The availability of subject-level data is an important consideration in deciding which studies to include in the meta-analysis. For reasons already discussed, subject-level data improve the quality of the meta-analysis by providing the ability to evaluate important quality factors of the component trials and possibly correct for any deficiencies identified, particularly poor outcome assessment or insufficient exposure periods. Subject-level data also allow for a broader range of analysis methods to be used and an examination of subgroups (see Section V). Note, however, that in some cases, meta-analyses based on only trial-level summary data may be able to identify or rule out risks associated with a drug. If so, then the criteria for determining which trials to include in a trial-level meta-analysis should be carefully considered; the principles described in this section apply to trial-level meta-analyses just as they do to subject-level meta-analyses.

I. Quality over Quantity

There is often a desire to include as many trials as possible in a meta-analysis to both increase the sample size and enhance the generalizability or external validity of the findings. Including trials that are of poor quality, however, does not accomplish this. The findings from a meta-analysis of a limited set of trials, selected with careful attention to trial and data quality, the intended use of the product, and combined using appropriate statistical methods, will yield a more informative answer to the safety question under investigation than a broader set of trials that includes trials of poor quality.

The criteria used to decide which of the candidate trials will be included in a safety meta-analysis should be carefully developed, taking into consideration outcome ascertainment and exposure periods as well as other factors described in the previous subsections. The choices of subject populations, dosing regimens, comparator arms, background therapy, standard of care, and other trial features that comprise the meta-analysis inclusion criteria will affect the validity and interpretation of the meta-analysis findings. Broad inclusion criteria (such as including trials where outcomes may not be reliably assessed) will likely compromise the internal validity of the meta-analysis without necessarily improving the external validity. The criteria for trial inclusion should be well-documented in advance of conducting the meta-analysis. This topic is discussed in section IV.

Trial inclusion decisions are particularly important for network meta-analyses, which are designed to assess safety concerns about one drug relative to another, when the two may not have been studied in the same randomized trial (Ohlssen, Price et al. 2014). Direct comparisons between drugs within the individual trials included in a network meta-analysis are used to form indirect comparisons between the two drugs of interest. Because some of the subject group comparisons are made across trials, it is important that the trials involved in a network meta-analysis be similar in design, subject populations, outcome definitions, and medical practice. Although the principles in this guidance apply to network meta-analyses, network meta-analyses have unique considerations beyond what is discussed in this guidance.

IV. THE IMPORTANCE OF PRESPECIFICATION AND TRANSPARENCY

The extent of the information that should be considered both before and following the conduct of a meta-analysis to adequately establish prespecification and transparency is discussed in this section.

A. Potential for Bias, Multiplicity, and Other Errors

Meta-analysis is a form of retrospective research in that most meta-analyses are conducted based on published clinical trials or trials already completed and whose results are known. It is important to minimize the potential for bias and other errors from sources that are often characteristic of retrospective research, including:

- Prior knowledge of individual study results when selecting the studies to be included in the meta-analysis
- Inclusion of the hypothesis-generating study in a meta-analysis designed to confirm the hypothesis
- Inability to determine the impact of multiplicity on the reported results

Special care is recommended when including trials whose results regarding the safety outcome of interest are known prior to the conduct of the meta-analysis. Information describing the knowledge base at the time the meta-analysis was planned will aid in determining the extent of possible bias that may affect interpretation of the results (e.g., trial outcomes influencing selection of trials). Prespecification of the criteria used to decide which trials to include before decisions about individual trials are made is a major mechanism to minimize bias and can help lessen the impact of this knowledge on the validity of the meta-analysis findings.

As stated earlier, our focus is on meta-analyses conducted to confirm a hypothesized safety risk. If a safety hypothesis was generated from the results of a specific clinical trial, then drawing inference from a meta-analysis that includes that trial is problematic. In this case, hypothesis test results and confidence intervals about the risk estimate are not readily interpretable. If the goal of the meta-analysis is to summarize existing information and not to make formal inference, then including the motivating trial may be reasonable. If the motivating trial is included, sensitivity analyses should be performed with and without the motivating trial to investigate its impact on the meta-analysis results (See Section V).

Another problem frequently encountered when evaluating the evidence provided by a meta-analysis is the potential for spurious findings due to multiple hypotheses being tested, multiple outcomes being evaluated, multiple or iterative analyses being conducted and multiple subject subgroups being investigated (Bender, Bunce et al. 2008). The result is inflation of the Type I error probability associated with the tests of hypotheses, making the meta-analysis conclusions difficult to interpret. When each of these sources of multiplicity is not well-described in advance, it is impossible to apply a statistical method of adjustment for multiplicity because the full range of factors that were evaluated cannot be determined. And even when the analysis plan does contain a clear description of the sequence of tests to be conducted (across hypotheses, outcomes, subgroups, etc.), there may be too little power available for each of the tests to

confirm the hypothesized safety signal, when appropriate adjustments are applied. Adequate planning and prespecification of meta-analysis objectives and tests of hypotheses may help minimize, to some extent, problems due to multiplicity.

B. Meta-Analysis Protocol

Prespecification, completeness, and transparency are important principles in the reporting of a meta-analysis, and the reporting begins with the meta-analysis protocol. The protocol should contain a detailed description of the information available prior to designing the meta-analysis that motivated the research. Potential problems anticipated in designing the meta-analysis and the methods planned to manage those problems should be documented. The protocol should be finalized prior to conducting the meta-analysis and, importantly, be in place prior to the selection of the component trials.

The meta-analysis protocol should be available through advance publication or other methods of distribution. This practice has been widely adopted for clinical trials via use of the web site, https://clinicaltrials.gov/. There are several repositories for the protocols, such as PROSPERO (Chien, Khan et al. 2012). Having protocols appear in the same publication as the meta-analysis findings is generally insufficient to provide such assurance.

Following is a list of the broad topics a meta-analysis protocol should include. Each is discussed further in the paragraphs that follow:

- The planned purpose of the meta-analysis
- The background information available at the time of protocol development that motivated the meta-analysis
- The design features of the meta-analysis, including outcome definition and ascertainment, exposure periods and assessment, comparator drugs, and target subject population
- A description of the search strategy that will be used to identify candidate trials and the criteria that will be applied for trial selection
- The analysis strategy for conducting the meta-analysis, including planned subgroup analyses and sensitivity analyses

Planned purpose: The planned purpose should be clearly stated in the protocol, with sufficient background material to explain the reason for conducting the meta-analysis. Examples include: to estimate a specific risk, to evaluate risk in a subgroup of patients, to identify risk factors or effect modifiers, to examine whether risk changes over time, or to assess accumulating evidence on product safety as ongoing studies of the product complete. The weight of evidence provided by a meta-analysis planned specifically to provide new information or update existing information about a hypothesized risk of a drug would be considered more compelling than that from a meta-analysis designed to explore safety signals or relationships among variables with no stated hypothesis. The distinction is analogous to that made between exploratory and confirmatory clinical trials in drug development, with the latter guided by pre-specified objectives reflected in a final protocol prior to study start.

Background information: The protocol should describe the information available prior to designing the meta-analysis that served as motivation for the research. Examples include safety risks identified in a randomized clinical trial of the drug or another drug in the same class, potential relationships between exposure and safety outcomes shown in post-marketing studies of health care data, or potential relationships identified during review of spontaneous adverse event reports.

Design elements: A clear prospective plan can help protect a meta-analysis against bias and inflation of Type I error by providing the rationale for each design element based on the knowledge and information available during planning. Without such a plan, it is difficult to determine which analyses were planned and which were exploratory or suggested as the analysis progressed. Important among the design elements is outcome ascertainment, including whether the outcome data were collected as part of the design of the individual trials or retrospectively collected as part of the meta-analysis; whether the outcome was actively collected from subjects or passively collected via subject adverse event reports; and whether the outcome was adjudicated, and, if so, how. Clear definitions of the outcome variable and the follow-up period for its ascertainment should be stated, with rationale for the choices thereof. The protocol should state the specific exposure of interest and the comparator. If multiple exposures (multiple doses of one drug or multiple drugs within a class) or comparators are to be combined, this should be stated, and the primary exposure and comparator should be identified.

Search and selection criteria: The protocol should describe the search algorithm that will be used to identify candidate trials to be considered for inclusion in the meta-analysis. Details should include a description of the sources to be searched, such as the literature or online resources (e.g., https://clinicaltrials.gov/, https://www.accessdata.fda.gov/scripts/cder/drugsatfda). The trial inclusion criteria should be described in detail, with the rationale given for each factor used as a basis for trial selection (see Section III.D). The selection process should be masked to study outcome and described in the meta-analysis protocol. Note that even if results are known to some parties, it may be possible to find others who could apply the trial selection criteria for the meta-analysis in an unbiased manner.

Analysis strategy: The protocol should describe the primary analysis strategy for achieving the study objectives as well as any sensitivity analyses and subgroup analyses planned. The statistical methods for the primary analysis should be stated in the protocol, with additional details provided in the statistical analysis plan. The analysis plan should be finalized prior to conducting the meta-analysis, analogous to the recommendation that a clinical trial's analysis plan be finalized prior to unmasking of treatment codes. Sensitivity analyses should be planned a priori to assess the impact of any unverifiable assumptions on the meta-analysis results. The factors that should be considered in choosing the statistical methods are discussed in section V.

C. Reporting Results from a Meta-Analysis

Results of a meta-analysis should be reported in a way that provides transparency and full disclosure of the many decisions involved in conducting the meta-analysis. The report should provide enough detail about the selection of trials, the statistical methods applied in the analyses, the results of those analyses, and the rationale for and results of any sensitivity analyses carried

out to enable an evaluation of the impact of bias and multiplicity on the findings and to assess their strength and credibility. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement provides some recommendations on the reporting of systematic reviews and meta-analyses (Moher, Liberati et al. 2009). Although not all the PRISMA components directly apply to meta-analyses that are the focus of this guidance, they should be considered.

The report should include the results of the search algorithm used to identify candidate trials and contain enough detail to evaluate the search. The selection process used to determine which of the candidate trials were selected for inclusion in the meta-analysis, which should be by applying the pre-specified criteria, should also be reported. Accounting for the trials that were not selected and the reasons for their exclusion is as important as accounting for the trials that were selected.

Characteristics of the individual trials included in the meta-analysis should be summarized, including individual trial design features, durations of exposure and follow-up periods, and patient populations. The report should describe the sources of any trial-level and subject-level data used in the meta-analysis. Summaries of subject-level characteristics should also be provided for the trials to be included in the meta-analysis, including basic demographics, concomitant medication usage, and other important factors thought to impact the exposure-risk relationship under investigation.

Any departures from the planned statistical methods should be described, as well as the rationale for those departures. Additional sensitivity analyses determined to be needed after the protocol was finalized, because of characteristics of the particular trials selected, unanticipated data issues encountered during analysis (e.g., zero-event trials), or preliminary findings needing further exploration, should be described and justified.

Results corresponding to the pre-specified test of hypotheses, supporting analyses, and sensitivity analyses should be provided in a clear and concise manner, with sufficient detail to aid in interpretation. Point estimates of absolute or relative risk should be accompanied by measures of uncertainty, e.g., confidence intervals. Forest plots are recommended for providing visual summaries of the results from each of the component trials relative to the results of the meta-analysis. These plots are useful in describing study-to-study heterogeneity.

V. STATISTICAL METHODOLOGY CONSIDERATIONS

A. Overview

In this section, general recommendations for selecting the statistical methods that will be used to combine evidence from the component trials in a safety meta-analysis are discussed. It is not the goal of this guidance to propose any best method, as no method performs best in all settings, nor is it the goal to restate the relative performance of methods that are well-established and have been compared in the literature pertinent to safety meta-analyses. Rather, this guidance recommends that the statistical methods used in a meta-analysis be aligned with the analysis objectives and hypotheses under investigation and be consistent with the study designs and data collected in the individual trials. The choice of methods should be justified based on the stated

objectives and documented in the protocol or analysis plan (see Section IV); sensitivity of the results to departures from assumptions required for correct application of the methods should be examined as part of the planned analysis strategy. Note that although this guidance generally recommends the use of subject-level data when available, it is recognized that meta-analyses conducted based on trial-level data only may be useful in certain settings (see Section VI.B.). The recommendations in this section apply to trial-level meta-analysis as well.

The material in this section falls into three broad areas: statistical properties of the analysis methods, heterogeneity, and sensitivity analysis.

B. Statistical Properties of Risk Estimates and Hypothesis Tests

The statistical approach for a safety meta-analysis should ensure that the estimator and/or hypothesis test have good statistical properties, namely that the resulting risk estimate is approximately unbiased and sufficiently precise, the standard error of the estimated risk is accurate, and the associated confidence intervals have accurate coverage properties. Tests of hypotheses about the risk should have good operating characteristics, i.e., the Type I and II error probabilities should be accurate, and the power maximized given the data available.

An important principle involved in estimating risk from a meta-analysis is that the randomized comparisons of the individual trials should be maintained when analyzing the combined data. In other words, when comparing drug A to drug B, subjects randomly assigned to drug A in a single trial are compared to subjects assigned to drug B from the same trial and not to subjects from other trials. In the statistics literature, this is referred to as stratifying the analysis by trial. Intuitively, this implies that the overall comparative measure of risk is based on combining the comparative risk measures from the individual trials using recommended statistical methods. Stratifying the analysis by trial is preferred to combining data across all subjects in the component trials by subject group prior to estimating risk, often referred to as simple pooling, as this ignores the randomized comparisons of the individual trials and can produce misleading findings.

When one or more of the trials included in the pooling does not employ a one-to-one randomization scheme, simple pooling of trial data can result in a phenomenon known as *Simpson's paradox* (Chuang-Stein and Beltangady 2011). When there are large sample size disparities among the trials with different randomization allocations, the impact of this phenomenon can be quite large. The hypothetical example in Table 1 illustrates an extreme example of Simpson's paradox in which, for each of four trials, the estimated risk of a safety event is identical for both Drug A and Drug B. With simple pooling, however, the risk for Drug A appears to be more than twice as high as that for Drug B (12.8 percent vs. 6.2 percent).

Table 1. An Illustration of Simpson's Paradox from Incorrect Pooling of Data

	Drug A			*Drug B*		
Trial	Events	Patients	Risk	Events	Patients	Risk
1	1	100	1.0%	2	200	1.0%
2	1	100	1.0%	2	200	1.0%
3	200	1200	16.7%	50	300	16.7%
4	2	200	1.0%	2	200	1.0%
Total	204	1600	12.8%	56	900	6.2%

It is sometimes of interest to combine multiple doses of a drug in one or more of the component trials to gain statistical power and improve the precision of the risk estimate in a meta-analysis. The combination of arms should be performed within each trial and the overall analysis should still be stratified by trial to avoid Simpson's paradox in this setting.

Sparse data resulting from rare safety outcomes pose particular problems in a meta-analysis. The statistical methods chosen for the analysis should perform well when the number of outcome events is very small in one or more of the component trials or in one or more treatment groups within a trial. Some commonly used methods perform well when there are ample events, but not so well when events are sparse (Bradburn, Deeks et al. 2007). For example, inverse variance weighting involves estimating risk with a weighted estimate of trial results, where weights are computed as the inverse of the trial level variance estimates. With sparse data, the estimated variances may not be well-determined, resulting in an unstable risk estimate. If some of the component trials have no events, the choice of methods is even more limited.

We do not recommend the use of continuity corrections, one approach for handling zero-event trials or trials with zero events in one or more treatment groups. Because it may not be apparent with some software packages if and how continuity corrections are incorporated, caution is needed to avoid their inadvertent use. Continuity corrections approaches generally involve adding small quantities to the zero event counts prior to analysis. Although their use allows zero-event trials to be included in a meta-analysis, the results may be biased. Note that the use of ratio effect measures, such as the risk ratio or hazard ratio, is more challenging in the presence of zero-event trials than is the use of risk difference measures, such as the Mantel-Haenszel risk difference (Greenland and Robins 1985). Another approach is to consider Bayesian methods for meta-analysis (Sutton and Abrams 2001) (Spiegelhalter, Abrams et al. 2004), which can incorporate information on trials with no events, even when a relative risk measure is used. The performance of any proposed method for dealing with zero-event trials should be established and the choice justified for a particular meta-analysis application.

The ability to replicate the results of a meta-analysis with an independent study will increase the persuasiveness of the findings. One such approach is to analyze one or more newly available trials to see if the results agree quantitatively and/or qualitatively with the results of an existing meta-analysis. Alternatively, an existing meta-analysis can be updated as new trials become available. Although this sequential approach to meta-analysis provides an efficient way to update risk estimates with new study results, the impact of repeated hypothesis tests about that risk should be taken into account (Whitehead 1997).

C. Heterogeneity

In any meta-analysis, heterogeneity of the drug effect among the component trials is expected and should be addressed at the design stage. If there is strong reason to believe trials will have importantly different drug effects based on known factors such as characteristics of the trial populations, the specific interventions, or other trial design features, then the statistical analysis should account for this expected variation. This may involve use of a statistical model that allows for different effects based on known factors. Alternatively, it may be of interest to conduct separate analyses for distinct groups of trials that vary with respect to one or more important design factors. For example, if the set of component trials consists of both placebo-controlled and active-controlled trials, a reasonable approach would be to perform a meta-analysis for each group of trials separately, taking into account what is known about the active control effect. In some situations the trials may be so heterogeneous that it is not possible to conduct a meta-analysis.

The most common approach to account for residual heterogeneity in drug effects across trials, after accounting for expected heterogeneity attributable to known factors, is to incorporate individual-trial treatment effects in the analysis model as either fixed or random effects. The meta-analysis literature includes a great deal of discussion about choosing between the two (Borenstein, Hedges et al. 2010). In the context of a meta-analysis, use of a fixed effects model is often interpreted as assuming a common effect exists across the trials, in contrast to the use of a random effects model, where the effects are assumed to vary across trials according to some probability distribution. This distinction is not usually made in other, similar areas of application, e.g., in managing centers in a multi-center trial (Senn 2000). In the statistics literature on multi-center trials (see, e.g., ICH E9), use of a fixed effects model is not as restrictive in that the model can specify either a common effect across centers or different, but non-random, effects for each center (i.e., by including the center by treatment interaction terms in the model). In the latter case, interest lies in estimating an average effect across the centers. Similarly, in meta-analysis, it may be desirable to allow effects to vary by trial with the inclusion of treatment by trial interaction terms in the fixed effects model, and, in this case, averaging across trials with appropriate methods provides the drug effect of interest.

Use of a random effects model in a meta-analysis implies an interest in estimating the average effect for some larger population of trials that are believed to be adequately represented by the trials in the analysis, and this parallels use of a random effects model in a multi-center trial; i.e., interest lies in estimating the average effect for a larger population of centers for which the trial's centers provide adequate representation. Arguments may be made against the use of random effects models in a meta-analysis based on the belief that the trials available for analysis are not a random sample of some larger population of trials — that is, all relevant trials are included in the meta-analysis. It has been pointed out, however, that even when there is no interest in making inference to a larger population of trials, use of a random effects model may produce more appropriate results, due to the better characterization of the between- and within-trial variance in the estimation process (Permutt 2003).

Both frequentist and Bayesian methods are available for random-effects meta-analysis, and the difference between the two lies in the assumptions made about the distributions of the random

effects, with Bayesian methods offering more flexibility (Muthukumarana and Tiwari 2016). Bayesian methods also allow multiple sources of variation to be incorporated in the modeling and estimation process. For example, in a meta-analysis designed to examine a specific risk for a class of drugs, one may assume there is a component of variation among different drugs within the class and a separate component among trials involving a single drug. To date, the Agency has limited experience in evaluating meta-analysis submissions that use Bayesian methods, but supports the consideration of Bayesian and other methods that achieve the desired properties discussed in this section.

For safety meta-analysis, the goal is to determine whether a significant risk is causally related to exposure to the drug, and the power available for that test should be maximized. Use of a fixed effects model will usually provide optimal power for detection of risk and also reflects a primary interest in the average effect among only those trials included in the meta-analysis. The parallel with the establishment of efficacy for drug approval is relevant here. The selective populations included in premarket efficacy trials may not fully represent the broader patient populations seen in clinical practice, but are still central in making regulatory decisions. However, for the quantification of the risk itself, a random effects model might be more appropriate, as the incorporation of the between-trial variance might better reflect the uncertainty associated with the risk estimate. Under all scenarios, the statistical inference should properly reflect the assumptions made for the fixed or random effects model used; in particular, the variance of the estimator should properly reflect whether the trial effects are constant, non-constant, or random.

D. Sensitivity Analysis

Sensitivity analyses play an important role in examining the impact of meta-analysis design decisions on the findings as well as the strength of evidence provided by the meta-analysis. The goal of any sensitivity analysis should not be to search for additional findings, but to support and understand the primary findings of the meta-analysis. Trial inclusion criteria, outcome definition, time period within which the safety outcome of interest is to be measured, and analysis method are examples of design characteristics that may be varied as part of a sensitivity analyses.

For example, if a meta-analysis protocol and statistical analysis plan called for including only those safety events that occurred during exposure periods in the risk estimate, then a sensitivity analysis that included all reported events, regardless of whether subjects were on or off drug, could provide important information about the observed risk estimate. A decreased event rate in off-treatment periods could, in this example, support causality (depending on the hypothesized mechanism). Similarly, a meta-analysis that included one very large study contributing a large proportion of subjects and events could raise a concern that it was overly influencing the meta-analytic results. A sensitivity analysis that excluded that study would have reduced numbers of subjects and events and lower power to yield a significant finding, but a risk estimate that was consistent with the original estimate would add to the weight of evidence of the finding.

It is often of interest to examine the consistency of findings from a meta-analysis across subgroups based either on trial-level or subject-level characteristics. Trial-level factors that might be of interest include the comparator treatment, dose and duration of treatment, background therapy, and subject inclusion criteria. Subject-level factors may vary within trials, and subject-

level data are required to provide estimates for each subgroup. In the antidepressant meta-analysis example of section VII, age was an important factor of specific interest, because the meta-analysis was motivated by an earlier meta-analysis of pediatric subjects. The number of subgroups to be examined should be kept to a minimum to avoid the consequences of multiple testing. Given the multiplicity issues, subgroup findings are seldom viewed as definitive in safety meta-analyses.

VI. STRENGTH OF EVIDENCE AND REGULATORY DECISIONS

A. Critical Factors in Determining the Strength of Evidence

Regulatory decisions related to drug safety are generally taken after considering the totality of available evidence, which may include meta-analytic findings, as well as other factors such as risk-benefit considerations, availability of alternative treatments, biological and clinical plausibility of the drug-risk relationship, and available regulatory options. The strength of evidence provided by the meta-analysis may influence a regulatory decision by FDA. The factors discussed above that FDA generally considers in determining the strength of evidence with respect to a safety-related regulatory decision can be summarized as follows:

- Quality and appropriateness of the individual trials for the meta-analysis objectives
  - Quality and completeness of safety outcome ascertainment
  - Appropriateness of studied populations and exposure and follow-up periods
  - Protocol adherence in the individual trials (e.g., compliance with investigational treatment, loss to follow-up, etc.)
  - Availability and quality of subject-level data
- Prespecification and adequacy of documentation
  - Prespecification and documentation of objectives, available knowledge, trial inclusion criteria, and choice of comparators, outcomes, statistical methods, and subgroups
  - Documentation that trial outcomes were not used as part of the trial selection criteria
  - Documentation of meta-analysis results including summaries of trials, subjects, outcomes, effect estimates, measures of uncertainty, and sensitivity analyses
- Appropriateness of statistical methods
  - Approach used for combining trials
  - Methods to handle sparse data or rare events
  - Methods to address heterogeneity
  - Sensitivity analyses
  - Validity of uncertainty estimates (e.g., confidence intervals or credible intervals)

Although not previously discussed, the magnitude of the estimated risk and associated measures of uncertainty are also important. A large estimated risk will generally be more convincing than a small to moderate one, because it will provide more assurance that an effect is real even in the presence of potential biases. Similarly, smaller p-values or narrower confidence intervals, both measures of uncertainty, provide additional assurance on the findings of the meta-analysis. For

safety meta-analyses, however, there is potential for bias from both known (e.g., selection of trials based on their outcomes) and unknown (biases that cannot be identified from the data used to conduct the meta-analysis) sources. Given this difficulty, standard measures of uncertainty, such as significance levels, should be interpreted with caution.

One approach to account for the many potential sources of bias and error in a meta-analysis is to replace the commonly used test size or alpha level for hypothesis testing, $\alpha = 0.05$, with an arbitrarily lower value (e.g., 0.01 or 0.001) in order for the results to be considered convincing. The choice of a lower value would reflect the recommendation to compensate for known and unknown sources of potential bias as well as to minimize the impact of multiplicity resulting from multiple comparisons. Such an approach would be important if the meta-analysis is the only basis for decision-making, as it will explicitly reflect the higher degree of uncertainty that exists for meta-analysis results. At the same time, there are often other sources of safety information so that the significance level for the meta-analysis is only one of many factors taken into consideration. Consequently, no single test size (alpha level) and no single confidence level can be recommended for deciding the level of statistical significance for results from a safety meta-analysis to be relied upon. The potential for harm may be so serious that marginally significant findings could prompt regulatory consideration. In this setting, however, the sources of bias and error related to the meta-analysis should be identified and accounted for wherever possible.

In addition to the magnitude of the observed effect and the level of uncertainty, the robustness of the risk estimate to appropriate sensitivity analyses can also support the strength of the meta-analysis findings. The importance of sensitivity analyses is described in section V, as their results play an important role in determining the strength of evidence. Risk estimates that are reasonably robust to the inclusion or exclusion of particular studies, or to changes in the statistical analysis methods used and assumptions required for appropriate use of those methods, will carry a greater weight of evidence than estimates that vary widely with such changes.

Similarly, risk estimates that are consistent across trials will also carry greater weight. In section IV, the use of forest plots or other graphical display of the study-specific risk estimates and their confidence intervals is advocated as a descriptive assessment of study-to-study heterogeneity. Absent any known cross-study differences, a high degree of similarity among study-specific results will strengthen the evidence provided by the meta-analytic summary risk estimate. Conversely, a large amount of variability among studies would make a marginal risk estimate (in terms of lack of statistical significance or small in magnitude) less persuasive.

B. Hierarchy of Evidence for Decision-Making

The factors described above for evaluating the strength of meta-analytic findings can be used to define a hierarchy of evidence against which meta-analyses conducted or reviewed for regulatory purposes should be evaluated.

- A top tier meta-analysis is one that is prospectively planned prior to the conduct of the trials to be included, and where the component trials are designed with the meta-analysis objectives in mind. The trials have well-ascertained outcomes and exposure periods, and

subject-level data are available for analysis. This level represents a gold standard not often realized in practice but useful as a benchmark in evaluating the quality of a meta-analysis.

- The next level down is a prospectively planned meta-analysis based on existing trials that were designed for other purposes but for which the quality of the data and the ascertainment of outcomes and exposure are adequate to support the planned analysis. Further, all meta-analytic study plans and trial inclusion decisions were made without knowledge of the study outcomes for the safety events of interest.
- The lowest tier, representing the least useful evidence for regulatory decision-making, corresponds to meta-analyses for which prospective planning did not occur, or is in doubt, study outcomes and trial inclusion decisions were made with outcome data in hand, and one or more of the important quality factors is in question, e.g., lack of rigor in outcome ascertainment, lack of subject-level data for use in determining exposure, use of inappropriate statistical methods such as simple pooling of trial data, or other issues.

Between the bottom and top tiers lies a broad range of meta-analyses for which an evaluation of the strength of evidence provided should include careful consideration of the important factors delineated in the previous subsections.

The level of evidence from a meta-analysis that is based solely on study level summary data, either prospective or retrospective, is generally considered to be lower than one for which subject-level data are available, as the party conducting the meta-analysis has little ability to judge the quality or completeness of the data or the appropriateness of the analysis methods used. On the other hand, if the outcome is relatively judgment-free and well-ascertained (e.g. mortality or perhaps stroke rate), these meta-analyses may still play a role in regulatory decisions. A study-level meta-analysis could be used as a first step to determine whether a more resource intensive subject-level meta-analysis is needed, perhaps based on the same studies. A hybrid would be a combination of studies for which subject level data are available for a subset; the mix would determine where in the hierarchy such a meta-analysis should be placed. The recommendations laid out in this guidance for producing high-quality meta-analyses apply regardless of the level (subject- or trial-level) of analysis involved.

There are two categories of meta-analyses considered particularly problematic for the regulatory framework and worth mentioning here. The first includes meta-analyses reported in the literature with no prior publication or credible record of a protocol to guide the selection of studies or prespecification of study objectives and analysis strategy. This type of meta-analysis is likely insufficient for regulatory purposes, for the reasons outlined in section IV. Even if the studies included in the meta-analysis represent a reasonable subset of those available (as opposed to only published studies), without documentation of a prespecified plan for deciding which to include and identifying outcomes of interest, it is usually not possible to determine what was known at the time the studies were selected, what analysis methods were chosen, or how many different analyses were conducted, in what sequence, and for which study populations or subgroups. Evidence from such an analysis would generally be considered too weak to support regulatory decision-making without further confirmation of the findings.

The second category includes meta-analyses that are based solely on safety results appearing in the literature. Limiting the meta-analysis to studies whose results appeared in publications about the exposure-risk relationship can introduce publication bias. This well-known phenomenon arises from the concern that studies failing to find a significant association between drug use and risk are not published at the same frequency as studies that show an association, and even among those published, bias may occur due to a failure to include certain safety outcomes in the publication and failure to include studies that did not show the outcome sought (Chalmers, Levin et al. 1987). Further, the information contained in the publication for each study may be lacking in detail, and without access to subject-level data, it may not be possible to rule out bias or severe heterogeneity in the results. For example, even if the results for the safety outcome of interest are reported for each trial, the details of how events were defined, measured, or adjudicated in the trials may not be clear, and it may not be possible to determine if the safety events of interest occurred on or off drug. Subject-level data are typically not available in publications of completed trials, limiting the ability to resolve these issues.

In summary, a number of important factors should be involved in determining the credibility of evidence from a particular meta-analysis. These factors range from the knowledge about and documentation of eligible studies, both published and unpublished; the quality and relevance of the studies selected as well as the process and timing of selection; and the validity of the statistical analysis that supports the inferential conclusions and the strength of the findings, evaluated against sources of potential or real bias. Whether or not the findings of a meta-analysis influence regulatory decision-making will generally depend, in part, on the strength of evidence provided by the findings, as determined by a careful evaluation of the important factors described in this guidance.

VII. EXAMPLES

A. Example 1: Antidepressant Use and Suicidal Events in Adults

This example illustrates the use of a meta-analysis to evaluate risks associated with a class of drugs and represents a prospectively planned meta-analysis of retrospective data, falling into the middle tier of the hierarchy of evidence discussed in section VI.B. The research hypotheses, study inclusion criteria, outcome measures, and statistical analysis plan were all specified prior to the conduct of the meta-analysis. Outcomes were uniformly adjudicated across studies, pooling of study data was accomplished with stratification, and subject-level data were available to explore subgroups as well as trends in risks across time. The interpretation of the findings, which resulted in a boxed warning for labeling of drugs in the class, reflects appropriate consideration given to the level of statistical significance, and to the consistency of findings.

In 2004, FDA completed a meta-analysis of studies of pediatric patients that showed an association of antidepressant drugs and suicidal behavior and ideation (Hammad, Laughren et al. (2006). Unsolicited information provided by drug sponsors to FDA and published articles motivated this meta-analysis. Based on the meta-analysis findings and deliberations from a

meeting in 2004 of FDA Advisory Committees (69 FR 47157[4]) and further consideration by FDA, a boxed warning was added to the labeling of all antidepressants concerning use in pediatric patients. Other FDA Advisory Committees later asked FDA to explore the association in adult patients (71 FR 66545[5].)

For this purpose, FDA planned and conducted a meta-analysis of randomized clinical trials of antidepressants. FDA is uniquely positioned to address this research question, because of the Agency's knowledge of marketed products in the drug class in question. The meta-analysis had several important features that supported its quality and utility for regulatory actions: (1) hypotheses generated from previous and independent evidence provided the meta-analysis objectives; (2) the meta-analysis was based on well-defined inclusion criteria and a complete set of the trials that met the inclusion criteria; (3) the meta-analysis employed rigorous and consistent outcome definitions across trials and patients; (4) the meta-analysis was based on a prespecified plan; and subject-level data was available.

FDA requested from all manufacturers of antidepressants all available subject-level data from randomized placebo-controlled trials of antidepressants. Basing the meta-analysis on data available to sponsors, while not inclusive of all potentially available data, has some important advantages. Because of regulatory requirements, trials from drug manufacturers typically contain detailed subject-level data including medical history, baseline characteristics, subject dispositions, patient outcomes, and adverse events. Focusing on the relatively small group of drug manufacturers (nine) allowed for the timely acquisition of the large amounts of pertinent data. Overall, the FDA obtained subject-level data considered usable for 372 trials.

The meta-analysis was prospectively planned but was based on previously collected data. Because the specific outcomes of interest were not systematically collected and adjudicated during the conduct of the trials, FDA provided specific instructions to the individual sponsors to conduct a retrospective identification and adjudication of potential suicidal behavior and ideation events from the subject-level data. The outcome definition required that the suicidal behavior and ideation events occurred on randomized treatment or within one day of stopping the randomized treatment. Based on adverse event reporting, potential events were identified with a specified algorithm. Based on blinded narratives of the events, qualified personnel classified the events into specific outcomes including: completed suicide, attempted suicide, preparatory actions toward imminent suicidal behaviors, and suicidal ideation based on the Columbia Classification Algorithm for Suicide Assessment (Posner, Oquendo et al. 2007). The overall process resulted in outcome measures that were consistently and rigorously defined across trials and subjects.

The meta-analysis employed a prespecified plan that included the trial inclusion criteria, hypotheses, outcome definitions, analysis methods, sensitivity analyses, and subgroups. The primary analysis method incorporated stratification by trial and accounted for the sparse nature of the outcome events by using exact statistical methods for hypothesis testing. Sensitivity

[4] Briefing package: http://web.archive.org/web/20040911055410/https://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4065b1.htm

[5] Briefing package: https://wayback.archive-it.org/7993/20170405070114/https://www.fda.gov/ohrms/dockets/ac/06/briefing/2006-4272b1-01-FDA.pdf

analyses were planned to examine the possibility and consequences of the following: differential exposure time between the randomized drug arms; heterogeneity of the effect measure across the trials; and trials with no events. The subject-level data allowed for the examination of important subgroups, including subject age, and for the examination of changing risk over time.

FDA presented the meta-analysis findings to a 2006 meeting of FDA Advisory Committees (FDA 2006) and sought advice on the interpretation and possible regulatory actions based on the findings of the meta-analysis. The meta-analysis found that the overall association of antidepressant drugs and suicidal behavior and ideation was not statistically significant in adult subjects, in contrast to the FDA meta-analysis of pediatric subjects. However, the association was nearly statistically significant for young adults, and a clear pattern emerged with respect to patient age (see Figure 1). The result from the pediatric meta-analysis supported this trend.

Based on the totality of the evidence, including results from the meta-analysis, FDA requested that manufacturers update the boxed warning on all antidepressants to include the risk of suicidal behavior and ideation associated with antidepressants for young adult patients in addition to pediatric patients. The warning states that the effect was not seen in adults over the age of 24, and for adults aged 65 and older, there was a reduction in risk. It should be appreciated that the clear pattern observed with respect to age and not just statistical significance led to the warning.

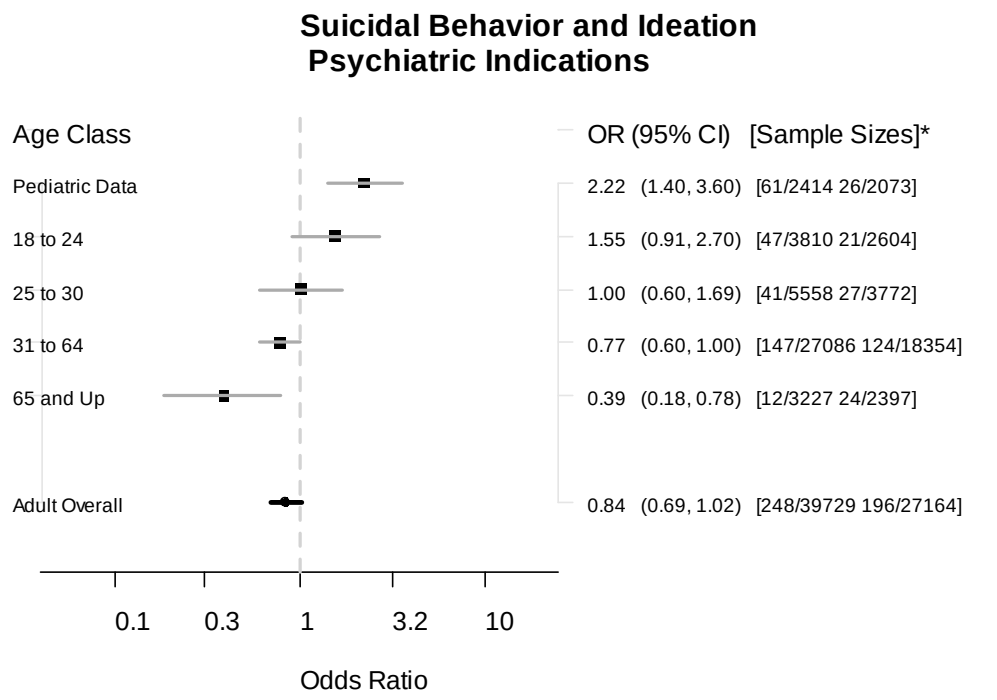

Figure 1: FDA Meta-Analysis of Antidepressants and Suicidal Behavior and Ideation.
Note: Pediatric results from previous FDA meta-analysis of pediatric patients (Hammad, Laughren et al. 2006).

B. Example 2: Tiotropium and Cardiovascular Events

This example illustrates FDA's consideration of the cardiovascular safety of the drug tiotropium and shows how the relative strengths of a well-designed, large, long-term trial and a meta-analysis based on published literature as well as other trial-level information (Michele, Pinheiro et al. 2010) were factored into a regulatory decision. Tiotropium bromide inhalation powder is a long-acting anticholinergic approved for use in treating bronchospasm associated with chronic obstructive pulmonary disease (COPD) and for reducing COPD exacerbations. The potential association of the drug with cardiovascular events was first reported to FDA based on an analysis of adverse events from 29 placebo-controlled trials conducted by the drug manufacturer. In particular, the simple pooled analysis of these studies showed that the drug had an excess number of strokes associated with its use. The pooled analysis was intended to identify potential signals for further evaluation and examine a range of adverse events. As is typical for such analyses, findings from the pooled analysis were not adjusted for multiplicity associated with examining multiple endpoints.

Because of the severity of the clinical outcomes, FDA issued a communication informing the public of the potential safety signal and FDA's efforts to investigate the findings. The communication noted that data from a large, four-year study called UPLIFT (Tashkin, Celli et al. 2008) would soon be available and would provide additional long-term safety data on the drug. Following the FDA communication, an article appeared on a meta-analysis of 17 randomized trials reporting a statistically significant increase in a cardiovascular composite outcome (cardiovascular death, myocardial infarction, and stroke) associated with inhaled anticholinergics (consisting of tiotropium and ipratropium) (Singh, Loke et al. 2008). At the same time, the results from UPLIFT had become available, and the initial review did not support a finding that tiotropium was associated with an increased risk of stroke, heart attack, or cardiovascular death.

A comparison between the pooled analysis of 29 trials conducted by the manufacturer and the UPLIFT trial highlights some important differences between the two sources of safety information. Although the pooled analysis contained more than twice as many subjects as UPLIFT (13,544 versus 5,992), the study duration of UPLIFT (4 years) was substantially longer than the durations of the trials in the pooled analysis (1 – 12 months). Consequently, UPLIFT provided more than twice as many person-years of follow-up (17,721 person-years) than the pooled analysis (7,636 person-years). Additionally, UPLIFT prospectively collected data on death and adjudicated cause of death for all subjects, including subjects who withdrew from the study.

In 2009, FDA convened an advisory committee meeting to discuss the results of UPLIFT and the published meta-analysis. The advisory committee concluded (11 votes to 1) that the results of UPLIFT adequately addressed the cardiovascular safety concerns that had been raised for tiotropium based on the initial pooling of 29 trials by the manufacturer and the published meta-analysis. The committee noted methodological concerns with the published meta-analysis, including lack of accounting for differential withdrawal rates between treatment groups and the potential for publication bias due to including only studies reporting an increase in cardiovascular events with use of tiotropium in the meta-analysis. The committee also noted concerns about the heterogeneity of trial designs in the review, including differences in study

drug, comparator drug, trial duration, and studied population. Based on the strength of the UPLIFT study findings and the methodological concerns of the published meta-analysis, FDA concluded that the available data did not support an association between the drug and adverse cardiovascular events.

The tiotropium example shows that a meta-analysis based on trial-level summaries may not agree with a large trial that is well designed specifically with a safety outcome as a primary objective. However, the Agency's position on the safety and effectiveness of a drug is based on the best information available at the time. In the tiotropium example, FDA issued a series of public communications to apprise the public of the latest safety information available and FDA's intended course of action. The example shows FDA's intention to carefully evaluate potential safety risks while balancing the need to not unnecessarily discourage or restrict the use of safe and effective drugs. The example also shows FDA's intention to act in a transparent manner, to the extent possible, based on the available data to ensure the safe use of drugs.

VIII. REFERENCES

Bender, R., C. Bunce, M. Clarke, S. Gates, S. Lange, N. L. Pace and K. Thorlund (2008). "Attention should be given to multiplicity issues in systematic reviews." J Clin Epidemiol **61**(9): 857-865.

Borenstein, M., L. V. Hedges, J. P. Higgins and H. R. Rothstein (2010). "A basic introduction to fixed-effect and random-effects models for meta-analysis." Res Synth Methods **1**(2): 97-111.

Bradburn, M. J., J. J. Deeks, J. A. Berlin and A. Russell Localio (2007). "Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events." Stat Med **26**(1): 53-77.

Chalmers, T. C., H. Levin, H. S. Sacks, D. Reitman, J. Berrier and R. Nagalingam (1987). "Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials." Stat Med **6**(3): 315-328.

Chien, P. F., K. S. Khan and D. Siassakos (2012). "Registration of systematic reviews: PROSPERO." BJOG **119**(8): 903-905.

Chuang-Stein, C. and M. Beltangady (2011). "Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies." Pharm Stat **10**(1): 3-7.

Greenland, S. and J. M. Robins (1985). "Estimation of a common effect parameter from sparse follow-up data." Biometrics **41**(1): 55-68.

Hammad, T. A., T. Laughren and J. Racoosin (2006). "Suicidality in pediatric patients treated with antidepressant drugs." Arch Gen Psychiatry **63**(3): 332-339.

Higgins, J. P. T., Green, S., (editors). (2011). Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

Michele, T. M., S. Pinheiro and S. Iyasu (2010). "The safety of tiotropium--the FDA's conclusions." N Engl J Med **363**(12): 1097-1099.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman and P. Group (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." PLoS Med **6**(7): e1000097.

Muthukumarana, S. and R. C. Tiwari (2016). "Meta-analysis using Dirichlet process." Stat Methods Med Res **25**(1): 352-365.

Ohlssen, D., K. L. Price, H. A. Xia, H. Hong, J. Kerman, H. Fu, G. Quartey, C. R. Heilmann, H. Ma and B. P. Carlin (2014). "Guidance on the implementation and reporting of a drug safety Bayesian network meta-analysis." Pharm Stat **13**(1): 55-70.

Permutt, T. (2003). "Probability models and computational models for ANOVA in multicenter clinical trials." J Biopharm Stat **13**(3): 495-505.

Posner, K., M. A. Oquendo, M. Gould, B. Stanley and M. Davies (2007). "Columbia Classification Algorithm of Suicide Assessment (C-CASA): classification of suicidal events in the FDA's pediatric suicidal risk analysis of antidepressants." Am J Psychiatry **164**(7): 1035-1043.

Senn, S. (2000). "The many modes of meta." Drug Information Journal **34**(2): 535-549.

Singh, S., Y. K. Loke and C. D. Furberg (2008). "Inhaled anticholinergics and risk of major adverse cardiovascular events in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis." JAMA **300**(12): 1439-1450.

Spiegelhalter, D. J., K. R. Abrams and J. P. Myles (2004). Bayesian approaches to clinical trials and health-care evaluation. Chichester, Wiley.

Sutton, A. J. and K. R. Abrams (2001). "Bayesian methods in meta-analysis and evidence synthesis." Stat Methods Med Res **10**(4): 277-303.

Tashkin, D. P., B. Celli, S. Senn, D. Burkhart, S. Kesten, S. Menjoge, M. Decramer and U. S. Investigators (2008). "A 4-year trial of tiotropium in chronic obstructive pulmonary disease." N Engl J Med **359**(15): 1543-1554.

Whitehead, A. (1997). "A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials." Stat Med **16**(24): 2901-2913.