DIA

Therapeutic Innovation
& Regulatory Science
2018, Vol. 52(2) 141-158

DOI: 10.1177/2168479017738980
tirs.sagepub.com

Sources of Safety Data and Statistical Strategies for Design and Analysis: Clinical Trials

Richard C. Zink, PhD[1,2], Olga Marchenko, PhD[3], Matilde Sanchez-Kam, PhD[4], Haijun Ma, PhD[5], and Qi Jiang, PhD[5]

Abstract

Background: There has been an increased emphasis on the proactive and comprehensive evaluation of safety endpoints to ensure patient well-being throughout the medical product life cycle. In fact, depending on the severity of the underlying disease, it is important to plan for a comprehensive safety evaluation at the start of any development program. Statisticians should be intimately involved in this process and contribute their expertise to study design, safety data collection, analysis, reporting (including data visualization), and interpretation. *Methods:* In this manuscript, we review the challenges associated with the analysis of safety endpoints and describe the safety data that are available to influence the design and analysis of premarket clinical trials. *Results:* We share our recommendations for the statistical and graphical methodologies necessary to appropriately analyze, report, and interpret safety outcomes, and we discuss the advantages and disadvantages of safety data obtained from clinical trials compared to other sources. *Conclusions:* Clinical trials are an important source of safety data that contribute to the totality of safety information available to generate evidence for regulators, sponsors, payers, physicians, and patients. This work is a result of the efforts of the American Statistical Association Biopharmaceutical Section Safety Working Group.

Keywords

adverse events, data monitoring committee, data visualization, multiplicity, meta-analysis, subgroup analysis

Introduction

Randomized clinical trials are the gold standard for evaluating the efficacy of any new intervention. Trials are adequately sized and powered so that for a small number of primary (and potentially secondary) endpoints, sponsors can establish the benefit of a novel therapy over the current standard of care in a predefined population of patients. Clinical development programs are designed to control variability and to ensure the quality of the generated data; therefore, the patients recruited to participate are those who meet a long list of study eligibility criteria. This naturally tends to exclude patients with other co-occurring disease, as well as those individuals taking 1 or more concomitant medications to address signs or symptoms, since both can potentially confound trial outcomes. Though efficacy is often the primary goal, sponsors collect myriad other data during the course of development to assess the safety of the intervention under investigation to better characterize the benefit-risk profile. For instance, phase I animal studies may indicate that the drug may not be safe for human use, phase I studies of cytotoxic chemotherapies in oncology patients may identify a maximum tolerated dose, or studies of dose-response in phase II could simultaneously assess efficacy and tolerability for selecting doses for further study. However, the examination of safety in early phases typically focuses on severe toxicities expected to occur in a small number of patients or healthy subjects over a limited duration of time. Further, animal studies may not be predictive of the effect observed in humans.[1] In practice, there are numerous safety endpoints to consider as a medical product proceeds through clinical development. Drug-related death and disease progression are obvious safety outcomes, and data for adverse events (AEs), laboratory abnormalities, vital signs, physical examinations, hospitalizations, electrocardiograms (ECGs), and patient-reported outcomes (PROs) for quality-of-life (QOL) can suggest other safety and tolerability concerns for the patient. In addition,

[1] JMP Life Sciences, SAS Institute, Cary, NC, USA
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[3] Bayer, Pharmaceutical Statistics, Whippany, NJ, USA
[4] SanchezKam, LLC, Alexandria, VA, USA
[5] Amgen, Thousand Oaks, CA, USA

Submitted 13-Jul-2017; accepted 25-Sep-2017

Corresponding Author:
Richard C. Zink, PhD, 701 SAS Campus Drive, Cary, NC 27513, USA.
Email: richard.zink@jmp.com

efficacy outcomes that fail to improve or worsen over time contribute to the questionable safety of the new treatment. The goal is to identify safety signals as early as possible, prevent or mitigate further safety issues where feasible, and to highlight areas requiring greater focus during postapproval safety monitoring.

There has been an increased emphasis on the proactive and comprehensive evaluation of safety endpoints to ensure patient well-being throughout the medical product life cycle.[2] In fact, depending on the severity of the underlying disease, it may be appropriate to plan safety assessments as early as the start of the development program. Though there has been some advancement in the quantitative evaluation of safety outcomes in clinical trials,[3-5] the methodologies available are often descriptive in nature because of a number of practical challenges:

1. Efficacy trials are sized and powered to identify differences among treatments for a small number of efficacy endpoints. Because of the rarity of many safety outcomes, the available sample size, even when combined across several trials, results in treatment comparisons for safety that are often underpowered.
2. Safety trials enriched with a sicker population to limit trial size and duration may create issues of generalizability.[6,7]
3. There are numerous safety endpoints that are repeatedly measured over time. If statistical testing is performed to identify differences among the study treatments without appropriate adjustment for multiplicity, this may lead to false positive findings. However, strict adjustment for multiplicity may further limit the already finite power available to detect safety signals. The analysis should strive for a reasonable balance of type I and type II errors.
4. Safety outcomes have important characteristics to consider including duration, severity, and investigator's assessment of causal relationship to drug, resulting in numerous sensitivity analyses.[8] Further, it is unclear which collection of event attributes would warrant consideration as the "primary" analysis.
5. Analyses within subgroups and by duration of therapy are important. For example, certain patient characteristics may contribute to an increased risk of safety outcomes, and certain patient populations, such as pediatrics, may have additional requirements to establish the safety profile.[9-12] Further, alternate estimands of patient safety should be considered. For example, safety outcomes are typically summarized for patients receiving at least 1 dose of study therapy, but analyses on alternate populations, such as patients who adhere to study therapy and/or complete the trial are important to consider.[13]
6. Not all safety endpoints and analyses can be prespecified. While the disease under investigation, the mechanism of action of study therapies, or animal models may suggest safety issues likely to occur during the course of the trial, unplanned safety issues may emerge, making it difficult to prespecify appropriate analyses in advance. For example, several drugs that have caused severe drug-induced liver injury (DILI) in humans have not exhibited clear hepatotoxicity in animals.[1]
7. Safety issues may occur spontaneously at any time during the trial, and often may occur between study visits. This adds complexity for summarizing results across time, and may result in some level of missing data for events that depending on the assumptions, could impact inference between study arms.
8. Medical classifications may be inaccurate and coded inconsistently.[2,14,15]
9. Safety could be associated with duration of therapy. Special consideration needs to be taken in studies where severe safety outcomes may lead to differential rates of drop out between the treatment arms (here we refer to the number of patients with the event per total number of patients), since patients with longer follow-up have greater opportunity to experience 1 or more safety outcomes.[16] The point estimates should always be interpreted in the context of the length of the observation/follow-up time. Note that drop out due to an adverse event is itself an important safety endpoint. Further, statistical challenges are present when the censoring mechanism is not independent of the event, as this is a common assumption in many time-to-event analyses.
10. All therapies carry some level of risk, and for more severe diseases, patients may be more willing to accept a greater degree of toxicity in order to obtain an important benefit than they would be for less grievous conditions. For example, natalizumab remains an important option for patients suffering from aggressive multiple sclerosis despite the potential risk of acquiring progressive multifocal leukoencephalopathy (PML), an opportunistic viral infection of the brain that can lead to severe disability or death. Patients with no prior immunosuppressant therapy and between 49 and 72 months of exposure to natalizumab have a 6/1000 rate of becoming positive for the JC virus, the underlying cause of PML.[17] Balancing the potential benefits and risks of new therapies is challenging, and this is currently an area of active research.[18,19]
11. Trials for chronic indications are too short to adequately assess long-term safety outcomes. Clinical trials are often short in duration compared to the treatment duration patients might experience in real life. Single-arm extension studies may be used to obtain long-term safety information, but this is at the expense of no concurrent control group to assess differences between the treatments (though open-label rates can

be compared to background disease rates; see Schnell and Ball[20] and Duke et al[21] for examples) . For chronic diseases, AE analyses in clinical trials will most likely underestimate the actual rate of occurrence and possibly even the severity of events. Additionally, "the safety evaluation during clinical drug development is not expected to characterize rare adverse events, for example, those occurring in less than 1 in 1000 patients."[22]

12. Safety analyses need to consider individual as well as collective ethics. Generally, in order to consider a therapy effective, a statistically and clinically significant treatment effect needs to be observed between the treatment arms. It is insufficient to identify individuals cases where patients may derive benefit from the drug, unless underlying patient characteristics can be shown to contribute to the response, as is the case for trastuzumab for breast cancer patients who are positive for human epidermal growth factor receptor 2 (HER2+).[23] As described above in (1), it may be difficult to identify population shifts in important safety parameters within an individual trial. However, whether a shift in population occurs or not for a given safety parameter, it is still necessary to identify individual patients experiencing severe outcomes in order for them to receive appropriate care. For example, many drugs pulled from the market because of severe DILI caused death or liver transplantation at rates ≤ 1 in 10,000.[1] The EudraVigilance Expert Working Group maintains a list of important medical events (IMEs) for which it may be important to screen AE data for the presence of individual cases.[24,25]

Because of the reasons outlined above, the unfortunate irony is that when the understanding of safety is at its most critical, for example, as it is in cytotoxic chemotherapies in oncology, the insight is more difficult to obtain. Despite these challenges, it is important to proactively plan for a comprehensive safety evaluation and signal detection at the start of any development program, a plan that considers the underlying challenges of the disease, as well as the unique features of treatment and patient management. This manuscript is the first in a series of papers from the American Statistical Association (ASA) Biopharmaceutical Section Safety Working Group to examine various sources of safety data and the statistical strategies for appropriate design and analysis. Our goal here is to describe and discuss comprehensive safety assessment across the plethora of endpoints collected in clinical trials; we limit discussion of safety outcomes and the related analyses where the safety endpoints are considered primary.[6,7,26] We describe the data sources available to influence the design and analysis of clinical trials supporting medical product development. We summarize the current state of safety analysis and reporting in clinical trials, and share our recommendations for the statistical and graphical methodologies necessary to appropriately analyze, report, and interpret safety outcomes, referring to US and European regulatory and other international guidance documents on safety where appropriate. Finally, we discuss the advantages and disadvantages of safety data obtained from clinical trials compared to other sources.

Available Data Sources for Safety

Safety Data From Premarketing Clinical Trials

Safety data is continuously evaluated at all stages of drug development, and the amount of data available depends on the stage of development. The safety profile of an investigational drug is determined from the analysis of safety information obtained from nonclinical and clinical studies. Nonclinical development is a stage of research that begins before clinical trials can begin and during which important feasibility, iterative testing, and drug safety data are collected. The nonclinical safety assessment for marketing approval of a pharmaceutical usually includes pharmacology studies, general toxicity studies, toxicokinetic and nonclinical pharmacokinetic studies, reproduction toxicity studies, genotoxicity studies, and for drugs that have special cause for concern or are intended for a long duration of use, an assessment of carcinogenic potential.[27] The clinical phase of development involves the following: phase I studies determine safety and dosing in healthy volunteers (patients in oncology studies); phase II studies initially determine efficacy and safety in patients with the disease or condition; phase III studies determine safety and efficacy in sufficiently large number of patients with the disease or condition. The safety information from clinical trials collected include AEs, laboratory measurements, vital signs, ECGs, and other tests relevant to the indication being studied.

The aim of drug safety regulation is to protect trial participants and patients from severe safety outcomes through early detection and prevention. During the safety review of a new drug application, regulatory agencies critically examine whether a drug is safe for its intended use. They assess the adequacy of testing for safety and determine the significance of the AEs and their impact on the approvability of the drug. Finally, regulatory agencies determine the safety issues to be included in the product labeling should the drug be approved and decide whether additional safety studies are needed.

Randomized controlled trials are the gold standard of scientific testing for new drugs. Based on the premarketing safety database, the relevant AEs and the risk factors for those events are identified. Adverse events are usually categorized according to the following domains: seriousness, expectedness of the event, relatedness, intensity, incidence, duration, latency, and time to resolution. The relationship between drug exposure and AEs is also assessed. However, there are inherent limitations to what can be learned from the premarketing safety database based on clinical trials.

Data Standards and Medical Coding

Data standards and medical coding dictionaries are important in that they make it easier to integrate information on safety outcomes from multiple sources, including clinical trials, observational or postmarket studies. The coding of events such as AEs is a process of converting investigators' verbatim terms to standardized terms. In the early 1990s, the International Conference on Harmonisation (ICH) identified the need for a standard international medical terminology for coding events for use in clinical trials worldwide. The adoption of a dedicated single standardized terminology offered a number of clear advantages for regulators, industry, and other stakeholders:

- Eliminated the need to convert data from one terminology to another, thus preventing the loss and/or distortion of data, and allowing for a savings in resources
- Improved the ease, quality, and timeliness of the data available for effective analysis, exchange and decision making
- Allowed for effective cross-references and analysis of data through a consistent terminology throughout the different stages of the development of a medicinal product
- Facilitated the electronic exchange of data relating to medicinal products

The Medical Dictionary for Regulatory Activities (MedDRA) is a medical terminology dictionary developed to classify AE information associated with the use of pharmaceutical and other medical products.[14] MedDRA employs a 5-level hierarchy of terminology allowing for the grouping of similar events, listed here in larger to smaller groups: System Organ Class (SOC), High Level Group Term (HLGT), High Level Term (HLT), Preferred Term (PT), and Lowest Level Term (LLT). It is used in the registration, documentation and safety monitoring of products through all phases of the development cycle (ie, from clinical trials to postmarketing surveillance). Regulatory authorities and the pharmaceutical industry are able to readily exchange and analyze safety data by coding adverse event data to a standard set of MedDRA terms. In addition to individual MedDRA terms, standardized MedDRA queries (SMQs) were developed to support signal detection and monitoring.[15] SMQs are validated, standard sets of MedDRA terms that have undergone extensive review, testing, analysis, and expert discussion by a working group of MedDRA and product safety experts. Some SMQs are a straightforward collection of terms, while others are based upon an algorithm of terms from more than 1 group. Some SMQs have hierarchical relationships with other SMQs. While MedDRA makes it possible to standardize event terminology, this coding step needs to be performed with care as improper coding may lead to missed safety signals. Examples of coding problems include splitting similar AEs among several PTs, lumping different verbatim terms to the same preferred term, or the lack of an adequate term or definition.

Corresponding to MedDRA for adverse events, the World Health Organization's Drug Dictionary (WHO-DD) is used to identify concomitant medications.[28,29] The WHO-DD is organized based on Anatomical Therapeutic Chemical (ATC) classifications. Each medicinal product is classified according to the primary organ or system on which it acts and its chemical, pharmacological, and therapeutic properties.

In addition to the use of a standard medical terminology dictionaries, there has been a push to have standards in the acquisition, exchange, and submission of clinical research data. The Clinical Data Interchange Standards Consortium (CDISC) is an open, multidisciplinary, neutral, tax-exempt, nonprofit standards-developing organization formed to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare.[30-32] CDISC Standards specify how to structure the data; they do not specify what data should be collected or how to conduct clinical trials, assessments, or endpoints. The US Food and Drug Administration (FDA) and the Pharmaceuticals and Medical Devices Agency (PMDA) of Japan require the use of CDISC standards for regulatory submissions. The use of data standards makes it more straightforward for regulatory agencies to analyze clinical trial data for safety signals using standard reports, and even combine data across multiple sponsors to assess the safety across similar medical products.

Current Statistical Strategies and Analyses for Safety

Safety Analysis Plans

A proactive strategy is necessary to ensure an objective and systematic evaluation of safety endpoints. A Program-wide Safety Analysis Plan (PSAP) was recommended by the Safety Planning, Evaluation and Reporting Team to document the statistical aspects of safety during clinical development and postmarketing activities (SPERT).[2] Although a PSAP is currently not required by regulatory agencies, several sponsors have adopted and implemented a PSAP (or a similar document by a different name) to document the appropriate data to collect and analyses to perform in order to characterize the safety profile of the new therapy throughout the product life cycle. As added benefits, the PSAP can facilitate ongoing interactions with regulatory agencies regarding current safety strategies, and can aid in the evaluation of the benefit-risk profile of the new therapy in the postmarketing stage. For example, the PSAP may be helpful in order to reach an agreement with a regulatory agency on the definition of an event of interest. This can be done through submitting the full or relevant section of a PSAP. The contents of a PSAP are flexible, and can be amended as needed. With respect to timing, initial development of the PSAP would begin prior to phase II in order to complete the first version of the standard data collection plan, with input from different functional groups such as biostatistics, safety,

clinical, and regulatory.[2] In contrast to the PSAP, the Statistical Analysis Plan (iSAP) for the Summary of Clinical Safety (SCS) limits analyses to development activities completed prior to submission, and is often finalized while phase III studies are ongoing.

Analysis Population

As ICH E9 points out, in the absence of drop out, poor compliance, failed eligibility criteria, or missing data, the analysis population would be "self-evident."[33] However, in the real world the appropriate analysis population is often less clear. Intent-to-Treat (ITT, strictly all randomized patients as randomized) is preferred compared to Per Protocol (PP, such as patients who complete the study and are compliant) as the analysis population for efficacy endpoints in superiority clinical trials.[33-36] However, arguments have been made for both the ITT and PP populations as the primary analysis population for non-inferiority clinical trials, though the most appropriate choice may depend on numerous factors that may be challenging to prespecify.[36-39] The primary concern over the analysis population, particularly for non-inferiority trials, concerns the bias inherent in estimated treatment effects, its magnitude, and direction. Ideally, the choice of analysis population should minimize this bias, or provide conservative estimates of treatment effects considering the goals of the trial. Like non-inferiority, safety analyses experience similar disagreements over the most appropriate analysis population. Consolidated Standards of Reporting Trials (CONSORT) recommends using ITT for safety endpoints.[40] However, the Council for International Organizations of Medical Sciences (CIOMS) suggested that safety analyses using the ITT population may underestimate treatment differences between the groups (often cited as an advantage in superiority trials of efficacy endpoints).[41] Further, a recent PhRMA safety working group favored the recommendations of the CIOMS.[2] Similar to CIOMS and the PhRMA working group, we recommend that safety outcomes, at least for superiority trials, are analyzed for patients receiving at least 1 dose of study therapy as they were treated (Safety Population).[2,41] The appropriate population for non-inferiority trials, similar to the debate for efficacy endpoints, is less clear and would benefit from further research.

The CIOMS report does suggest other analysis populations, such as those patients who receive a prespecified number of doses, though even the authors admit that such analyses may be biased, with an unknown direction of bias. Other possibilities to assess the impact of analysis population include summaries of safety generated at an earlier time point to capture all available patients, or periodic assessments based on the duration of exposure. Draft European regulatory guidance makes a similar recommendation for this latter point specifically to account for the bias likely to occur owing to differential patient dropout between study arms.[16] However, any notable differences between the treatment arms, such as drop-out rates and reasons or medication compliance, should be examined to determine the extent of any bias on treatment estimates. Further, substantial differences between the ITT and Safety Populations may require additional sensitivity analyses for safety outcomes. Similar to analyses of efficacy for ITT and PP populations, comparable findings between the ITT and Safety populations will provide comfort in the study results. Forthcoming revisions to ICH E9 may suggest further estimands appropriate for the analysis of safety.[13] However, the bottom line for clinical trialists is to minimize the extent of dropout and missing data and to maximize the quality and compliance in any trial.

Safety Monitoring

Data and safety monitoring in clinical trials can be defined as a planned, ongoing process of reviewing the data collected in a clinical trial with the primary purpose of protecting the safety of trial participants, the credibility of the trial, and the validity of trial results.[42,43] In guidance for clinical trial sponsors on the Establishment and Operation of Clinical Trial Data Monitoring Committees, the US FDA defined a clinical trial Data Monitoring Committee (DMC) as a group of individuals with pertinent expertise that reviews on a regular basis accumulating data from 1 or more ongoing clinical trials.[44] They noted that a DMC is also known as a Data and Safety Monitoring Board (DSMB), or a Data and Safety Monitoring Committee (DSMC). An independent DMC is usually responsible for the data and safety monitoring of randomized phase II and phase III studies. Though it is possible to have an independent DMC for phase I trials, it is not a common practice.

Sometimes safety reviews occur with the same frequency as efficacy reviews, but safety reviews typically take place more frequently. For interim safety review of randomized trials, unblinded summary data and blinded individual data are often provided to the DMC. Blinded monitoring of individual safety data is usually performed by a study team or a study coordinator. In addition to the safety analysis, an independent DMC may request unblinded summaries of efficacy to evaluate the benefit-risk profile when the study under review is considered for early termination owing to a safety concern. Very often industry-sponsored clinical programs utilize the same independent DMC for all trials in a development program. Although DMC decisions are advisory, they should be taken seriously by company executives responsible for patient safety and study validity and integrity.

As a follow-up to a 2012 guidance for safety reporting requirements for investigational new drugs (INDs) and bioavailability and bioequivalence studies, a 2015 draft guidance for industry on Safety Assessment for IND Safety Reporting, the FDA recommends that sponsors use a Safety Assessment Committee (SAC).[45,46] While its use is in early stages of implementation, the SAC should oversee "the evolving safety profile of the investigational drug by evaluating, at appropriate intervals, the cumulative serious adverse events from all of the trials in the development program, as well as other available important safety information (e.g., findings from epidemiological

studies and from animal or in vitro testing) and performing unblinded comparisons of event rates in investigational and control groups, as needed, so the sponsor may meet its obligations under [21 CFR Section] 312.32(b) and (c)."[47] The SAC is distinct from a DMC. In most cases, an existing DMC will not be able to function as an SAC because the DMC may meet too infrequently and is often focused on a single trial, rather than the entire safety database. Further, DMC recommends to the sponsor when to modify or stop a study because the investigational drug is ineffective or reveals an important safety concern. In contrast, the role of the SAC is to review accumulating safety data to determine when to recommend that the sponsor should submit an IND safety report to the FDA and investigators participating in the clinical program. Despite their distinct roles, the SAC can participate in a discussion with a DMC whether the conduct of a specific study should be revised based on the currently available safety information.

Selected Monitoring Strategies

Discrete Monitoring

Here we focus on safety monitoring and assessment when safety endpoints are primary or co-primary endpoints in a clinical trial.[46] For a general overview of safety monitoring in the premarket setting see Fries et al.[48] One example of a composite (composed of multiple safety endpoints) primary endpoint is Major Adverse Cardiovascular Event (MACE) or MACE+ in Cardiovascular Outcome Trial (CVOT) in patients with type 2 diabetes mellitus (T2DM). MACE includes CV death, nonfatal myocardial infarction (MI) and nonfatal stroke events, while MACE+ usually includes MACE plus hospitalization for unstable angina. Guidance on the assessment of CV risk for non-insulin therapeutics for T2DM requires sponsors to rule out excess CV risk pre- and postmarket, and suggests ways of integrating data from pre- and postmarket evaluations to assess risk margins.[49] In particular, the guidance specified that the upper bound of the 2-sided 95% confidence interval for the estimated CV risk of an experimental treatment compared to a control arm should be less than 1.8 in the premarketing evaluation and less than 1.3 to meet the postmarketing requirement. The ASA Biopharmaceutical Section Safety Working (BIOP SWG) Group published a manuscript on the evaluation of CV safety for drugs used to treat T2DM in an adult population approved between 2002 and 2014 by the FDA.[6] The group discussed different strategies to address pre- and postmarketing requirements for the CV safety in patients with T2DM:

1. Strategy 1
    a. Stage 1: Meta-analysis of CV events observed in phase II and phase III trials
    b. Stage 2: CVOT
2. Strategy 2
    a. Stage 1: CVOT
    b. Stage 2: Meta-analysis of the Stage 1 CVOT and a separate Stage 2 CVOT
3. Strategy 3
    a. Stage 1: Interim analysis of an ongoing CVOT
    b. Stage 2: Analysis of the completed CVOT
4. Strategy 4
    a. Stage 1: Meta-analysis of CV events observed in phase II and phase III trials and interim results of an ongoing CVOT
    b. Stage 2: Analysis of the completed CVOT.

A second manuscript published by the BIOP SWG outlined statistical challenges encountered during the design and analysis stages of CVOTs and shared approaches to address these challenges.[7] For example, the treatment effect observed at multiple interim analyses is a source of multiplicity in need of adjustment. The premarketing requirement to demonstrate non-inferiority to a hazard ratio of 1.8 and the postmarketing requirement to demonstrate non-inferiority to a hazard ratio of 1.3 can be handled using 1 or more CVOTs in combination with adjudicated events obtained from phase II and III studies via meta-analysis (eg, Strategy 3 or Strategy 4). The pre- and postmarketing requirements to rule out excess CV risk can be tested sequentially, or concurrent testing using interim results from CVOTs can be employed. Group sequential boundaries that permit early stopping to claim non-inferiority in both the pre- and postmarketing settings can be implemented in these cases. However, there is not much difference between the sequential and concurrent testing when conservative spending functions like the O'Brien-Fleming (OBF)–type spending function are used.[50]

Although the primary objective of a CVOT is to demonstrate non-inferiority, some sponsors are interested in testing for superiority to see if the new treatment demonstrates a CV benefit compared to a control. There are 2 general approaches to do this. The first strategy sizes a CVOT for superiority at the design stage (eg, SAVOR CVOT with 16,500 patients). The second strategy initially designs the trial with sufficient power to test for non-inferiority, with an interim adaptation to increase the number of events and patients to test for superiority at the end of the trial if the interim result is highly promising (eg, EXAMINE CVOT with 5400 patients). If the interim result is not promising for superiority, the trial stops with a non-inferiority claim (if reached) using fewer patients, and completing sooner with less cost.

Continuous Monitoring

In some disease areas, toxicity monitoring might need to be done on a continuing basis (eg, in oncology with cytotoxic drugs). Ivanova et al summarized safety monitoring strategies in oncology trials. Phase I oncology trials are designed to assess the toxicity of novel therapies by identifying the MTD.[51] However, given the relatively small number of patients in phase I oncology trials, the recommended phase II dose can be imprecisely defined, leading to excessive toxicity in a phase II trial. Most phase II trials are designed to terminate a study early if

the treatment is not promising for further development, but it is equally important to have stopping rules for toxicity. Frequentist and Bayesian methods have been developed to evaluate both toxicity and efficacy as bivariate variables. Most proposed methods are 2-stage and range from equal weighting of the efficacy and toxicity responses, to designs with variable trade-offs between these 2 outcomes.[52-54] Several authors proposed Bayesian methods for the simultaneous monitoring of both efficacy and toxicity that allow for trade-offs between corresponding rates.[55-60]

A single interim analysis to assess safety and toxicity is often not sufficient; continuous monitoring allows for stopping at any point should the toxicity rate be unacceptably high. Continuous monitoring rules can be based on Frequentist or Bayesian methodologies. Among Frequentist approaches, the Pocock boundary is recommended since it allows stopping the trial for toxicity as early as possible. Alternatively, the sequential probability ratio test (SPRT) leads to a boundary very similar to Pocock for a given sample size and type I error rate.[51] Geller et al proposed a Bayesian stopping rule where the trial is stopped if the posterior probability of the dose limiting toxicity rate exceeding the ideal target rate (typically 0.20) is equal to or higher than a prespecified value τ.[61] The value of τ is often chosen to be 0.90, 0.95, or 0.98. In practice, if prior information about the toxicity rate is unavailable or deemed inappropriate or unreliable by the study team, the Pocock or the Bayesian boundary with a non-informative prior can be used. However, the Bayesian boundary is recommended if there is reliable prior information about toxicity. For clinical trials in oncology where toxicities can be both life-threatening and/or take time to develop, statistical designs that consider both stopping and enrollment rules are recommended.[62]

Reporting Safety Analyses

Initial Steps

Death and disease progression, while important indicators of patient safety, are often analyzed as primary efficacy endpoints in clinical trials. Because the strong control of type I error is well understood in these situations, even in the presence of 1 or more interim analyses, we avoid further discussion specific to these endpoints within this article. Instead, we refer readers to key works on the statistical monitoring of clinical trials,[63,64] and highlight recent references where safety outcomes serve as primary endpoints.[6,7,26] Here, we focus on the efficient reporting of the considerable volume of safety endpoints that are collected within a clinical trial, with a primary focus on adverse events. Given the challenges inherent to the analysis of safety that were outlined above, it should come as no surprise that clear insight is often out of reach. The traditional means of data summary, such as tables and listings, are often ineffective for communicating the story hidden within the data.[65] Data visualization is the key to effective communication of safety outcomes; we reinforce this point through several examples below.

Our rationale for the focus on adverse events is due to the fact that occurrences of clinically relevant changes in other safety endpoints are reported as AEs. For example, significant changes in the laboratory test alanine aminotransferase, an important indicator of liver health, can be represented by the adverse events "alanine aminotransferase abnormal," "alanine aminotransferase increased," or "alanine aminotransferase decreased" when using MedDRA. These categorizations naturally occur in clinical practice (eg lab measurements above some multiple of the upper limit of normal determined by age, gender, and disease severity), though the resulting loss of information tends to "disappoint" some statisticians.[66,67] Despite our limited mention of other safety endpoints, many of the recommendations and analysis strategies made throughout this manuscript still apply. We illustrate the various methodologies using data from clinical trials of patients with aneurysmal subarachnoid hemorrhage, type II diabetes mellitus (T2DM), or chronic myeloid leukemia.[68-70]

Adverse events that occur since the previous study visit are reported to the clinician by the patient or caregiver. Additional AEs may be identified by the clinician through in-clinic or laboratory assessments that have worsened since baseline. Details on the severity or toxicity grade (National Cancer Institute's Common Terminology Criteria of Adverse Events [NCI-CTCAE]), seriousness, outcome, and duration of the event, along with the action taken with study drug due to the event, and the investigator's opinion on the relationship to study medication are recorded. As described above, verbatim text is coded using MedDRA to maintain consistency in the reporting and grouping of AEs within and across studies and development programs. AEs are traditionally summarized by PTs and grouped by SOC in order of decreasing frequency of occurrence. Binary outcomes, such as whether a patient experienced a particular AE or not, are often reported using a risk difference ($\hat{p}_{tj} - \hat{p}_{cj}$), risk ratio ($\hat{p}_{tj}/\hat{p}_{cj}$), or odds ratio ($\hat{p}_{tj}(1-\hat{p}_{cj})/(1-\hat{p}_{tj})\hat{p}_{cj}$), where $\hat{p}_{ij}$ is the probability of experiencing event j of J possible AEs for treatment i.[71,72] Pros and cons for the various measures are discussed in Zhou et al.[72] Risk differences are presented in Figures 1 to 3.

Given the large number of potential comparisons of treatment arms for adverse events, Crowe and coauthors suggested a 3-tier approach for the analysis of AEs.[2] Preplanned hypothesis for tier I events, those AEs expected to occur or of considerable clinical relevance for the disease (which may overlap considerably with AEs of special interest), would typically not receive adjustment for multiple comparisons unless there were numerous tier I events to consider. Treatment comparisons for unexpected commonly occurring (4 or more patients in a single treatment arm) tier II events should consider adjustments for multiple comparisons. Tier III events (those not in tiers I or II) are rare and should be summarized in a listing. Appropriate multiplicity adjustment for tier I (if required) and tier II events should achieve a reasonable balance between committing type I errors without overly sacrificing the power to detect potential safety signals. The false discovery rate (FDR) provides a more

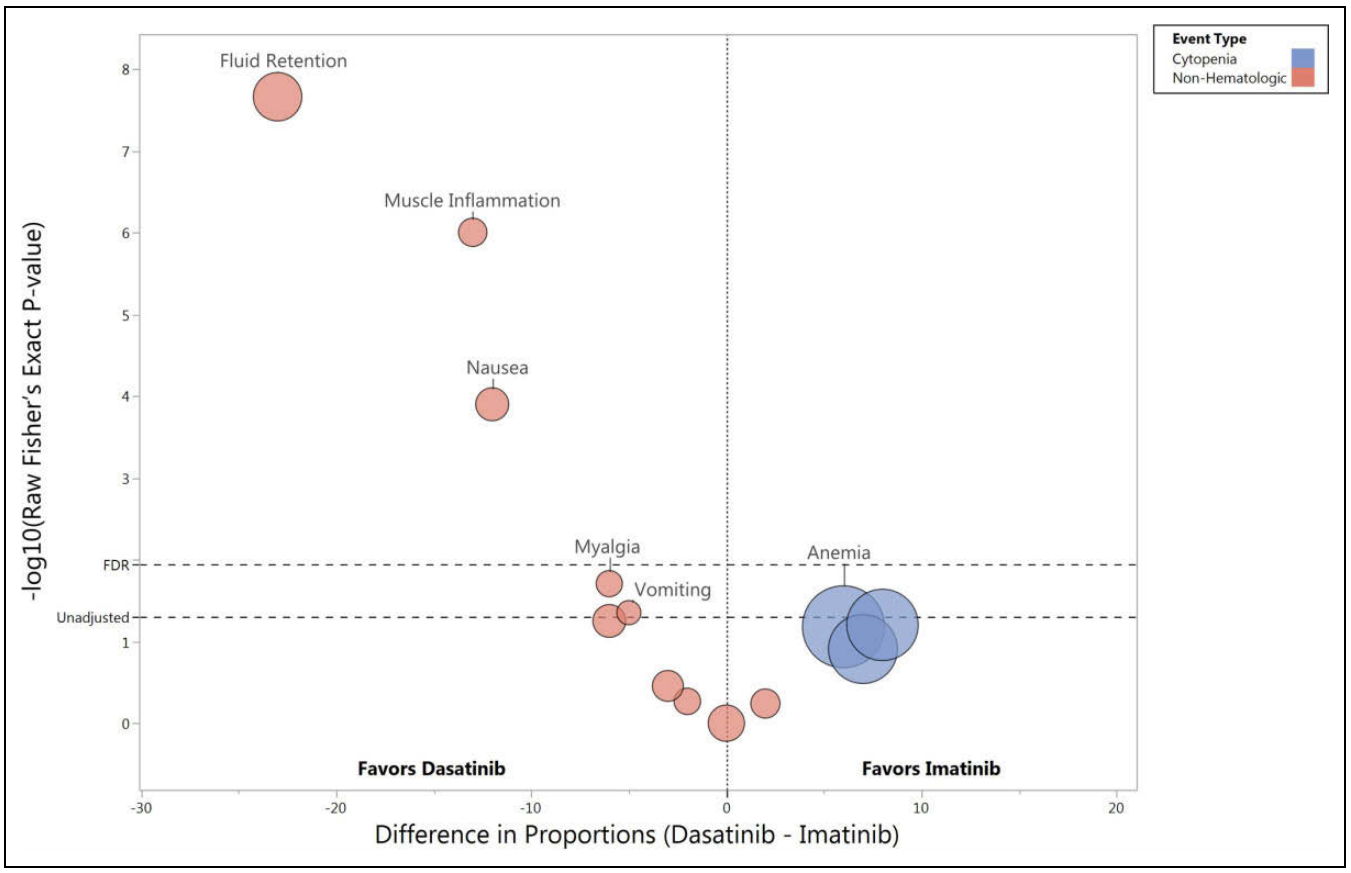

Figure 1. Volcano plot of drug-related adverse events that occurred in at least 10% of treated patients with chronic myeloid leukemia. Unadjusted vertical reference line drawn at $-\log_{10}(0.05) = 1.3$. FDR reference line drawn at $-\log_{10}(0.0115) = 1.9393$, where $\alpha^* = 3/13 \times 0.05 = 0.0115$. Horizontal reference line drawn at 0 to highlight no difference in risk between the treatments. Alternatively, the FDR reference line could be drawn at $-\log_{10}$(maximum unadjusted *P* value $\leq \alpha^*$) as in Zink et al.[77] The bubble area is proportional to the total number of patients that experience an adverse event for both treatments combined. Data from Table 4 of Kantarjian et al.[70] *P* values and confidence intervals were computed in Ivanova et al.[51] FDR, false discovery rate.

balanced approach between type I error and power, since it does not control the familywise error rate.[73] The FDR, typically prespecified at $\alpha = 0.05$, is the proportion of erroneous rejections among the rejected null hypotheses from a set of multiple tests. In general, with J treatment comparisons of ordered (smallest to largest) P values $p_{(j)}$, the FDR P value for the jth hypothesis is

$$p^*_{(j)} = \begin{cases} p_{(J)} & \text{for } j = J \\ \min\left(p^*_{(j)}, \dfrac{j}{(j-1)} p_{(j-1)}\right) & \text{for } j = 1, 2, \ldots (J-1) \end{cases}$$

Corresponding simultaneous 95% FDR confidence intervals can be defined by finding the largest j where $p_{(j)} \leq j\alpha/J$ and using $\alpha^* = j\alpha/J$ for all J confidence intervals.[74] An alternate FDR methodology that considers the relationship among AEs, the double FDR (DFDR), could also be considered for analysis.[75]

A volcano plot of tier I and II events (Figure 1) is used to summarize the incidence of adverse events for patients who receive at least 1 dose of study therapy, often referred to as the safety population.[8,76,77] Since this example is taken from the literature, we have no distinction between tier I and II events. However, the tiers could be distinguished in practice by varying the bubble pattern (solid vs striped), or providing separate plots for each tier. The x-axis represents the dasatinib minus imatinib risk difference while the y-axis represents the $-\log_{10}$ transformation of the unadjusted P value from Fisher exact test. Here, bubble area is proportional to the total number of all patients that experience the particular adverse event, while bubble color distinguishes event type. Alternatively, bubble area could be proportional to the inverse of the variance of the treatment effect. Unadjusted ($-\log_{10}(0.05) = 1.3$) or FDR-adjusted ($-\log_{10}(0.0115) = 1.9393$, where $\alpha^* = 3/13 \times 0.05 = 0.0115$) reference lines are drawn to emphasize statistically significant events, those events where the center of the bubble is above a particular reference line. Figure 1 clearly communicates that anemia and vomiting are the most and least common events; imatinib shows significantly greater FDR-adjusted risk for fluid retention, muscle inflammation, and nausea, with greater unadjusted risk for myalgia and vomiting; and that the events with greater risk (though not statistically so)

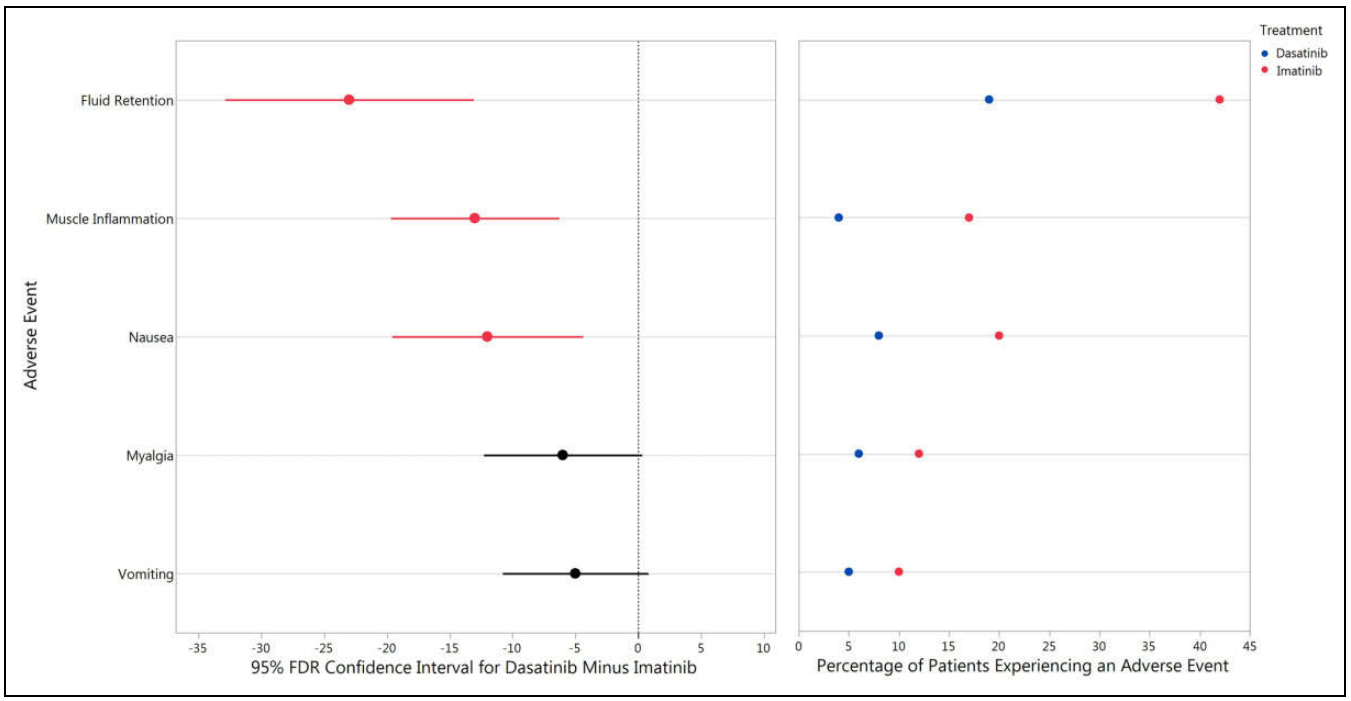

Figure 2. FDR confidence intervals and event incidence for identified safety signals in patients with chronic myeloid leukemia. Presentation suggested as in Amit et al.[78] Left panel displays a forest plot of FDR intervals for dasatinib minus imatinib for signals identified from Figure 1; red intervals indicate significantly increased risk for imatinib.[79] Reference line is drawn at 0 to indicate no difference between dasatinib and imatinib. Right panel presents a dot plot to communicate the incidence of each AE for each treatment arm. Data from Table 4 of Kantarjian et al.[70] *P* values and confidence intervals were computed in Ivanova et al.[51] FDR, false discovery rate.

on dasatinib are common and cytopenic. Such insight would be challenging to obtain from a table of summary statistics. To communicate additional details for important events identified from the volcano plot, Figure 2 summarizes FDR intervals and incidence rates using a forest plot and a dot plot.[78,79]

Before we proceed to the next section, it is worth noting that the above analysis in Figures 1 and 2 considers events that occur in at least 10% of treated patients, irrespective of severity, seriousness, or relationship to study therapy. In oncology, for example, similar displays are often presented for the subset of events viewed by the investigator to have a causal relationship with treatment, with additional summaries limited to those events with NCI-CTCAE grade 3 and above.[16] Bubble color within volcano plots can be used to illustrate different characteristics, for example, by coloring according to average severity or the proportion of events that are considered serious. Alternatively, bubbles could represent the cross-classification of preferred term with 1 or more covariates. However, this approach would likely limit power to identify meaningful differences between the treatments. Figure 1 can be reproduced considering the ordinality of varying characteristics to develop meaningful subsets of events, using a heat map to summarize the standardized effects across numerous volcano plots for patients receiving at least 1 dose of study medication. The standardized effect is the treatment difference divided by the standard error of the treatment difference (similar to a *Z*-statistic). The goal is to incorporate some level of statistical significance into the summary of the treatment effects, since the statistical significance for differences in proportions is dependent on the original proportions of the individual groups. In other words, the significance of a 10% difference will vary whether the original rates were 15% versus 5%, compared to 55% versus 45%. For example, Figure 3 shows elevated risk for placebo of vasoconstriction for all events, SAEs, and more severe events, while nicardipine has elevated risk for phlebitis across all event subsets, though with weaker results among more severe events. Heat maps, such as Figure 3, can be extended to assess the sensitivity of varying estimands and/or patient populations, such as those patients that adhere to study therapy.[13] Depending on the number of events, however, heat maps may need to be printed by system organ class, event tier, or limited to the most common events.

Further Analysis Issues

The previous section highlights a starting point for safety evaluation, and provides some recommendations for assessing how various event characteristics can impact analysis findings. However, there are several other event features to consider in order to adequately portray the story hidden within the safety data. The influence of time, which is related to the exposure to study intervention or time since surgery or another procedure, tends to be ignored in most presentations of AEs. Since patients with longer follow-up have greater

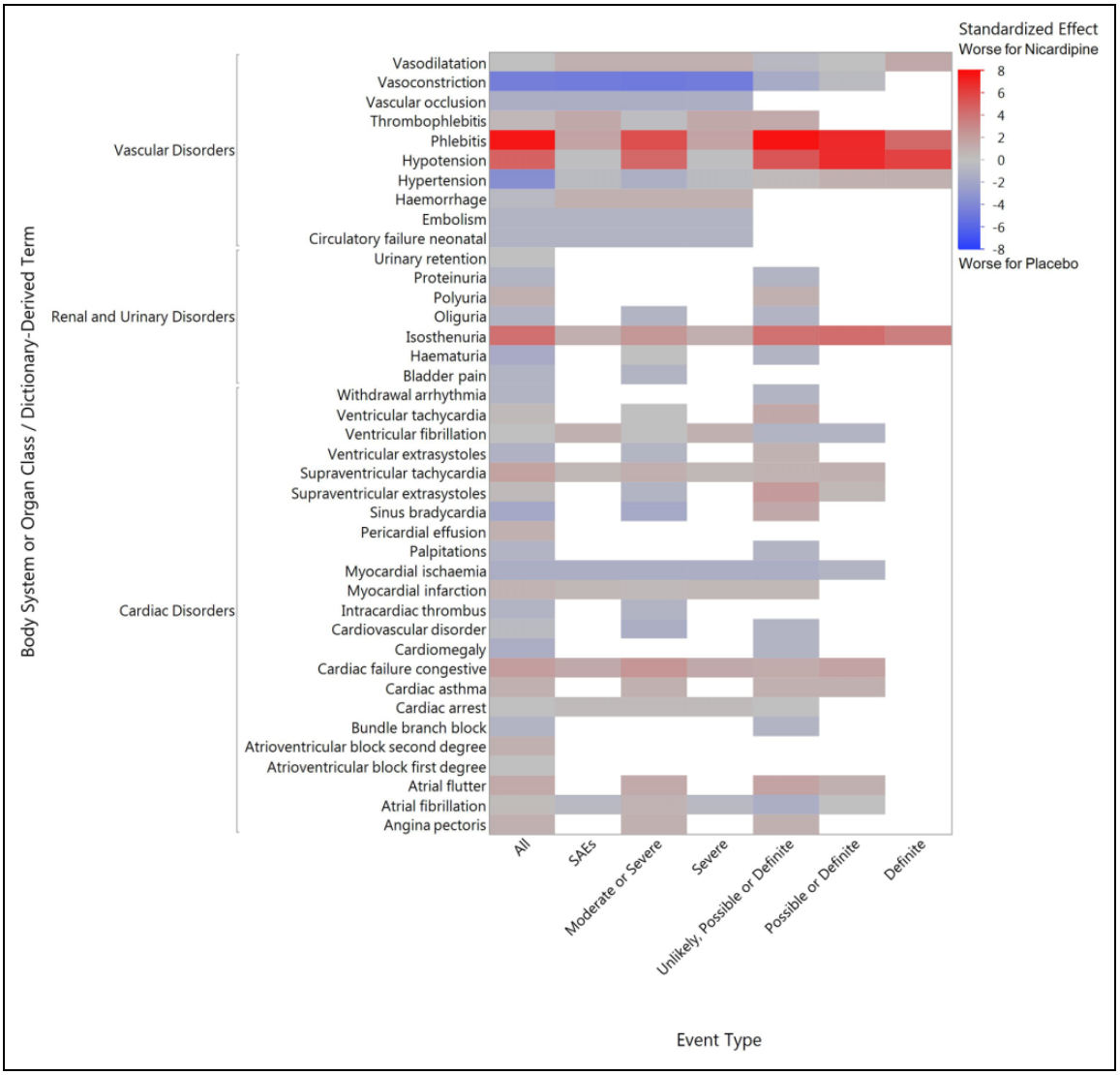

Figure 3. Sensitivity analysis of treatment emergent adverse events in patients experiencing a aneurysmal subarachnoid hemorrhage. Standardized effect is the risk difference for experiencing an adverse event for nicardipine minus placebo divided by its standard error. Darker red or blue indicates higher risk on nicardipine or placebo, respectively. Cells are white when the standardized effect cannot be calculated, most often when no events occur. Because of space limitations, a subset of system organ classes is presented. Data are from Haley et al.[68]

opportunity to experience 1 or more safety outcomes, it is important to consider exposure-adjusted incidence rates or time to first events in studies with varying patient exposure.[16,72,80-85] However, exposure-adjusted incidence rates assume a constant hazard rate across time. Liu and coauthors suggest that this assumption is likely to hold for rare events, though it should be assessed in practice since this expectation may not apply for many events.[81]

Though some additional assumptions may be required, analyses by time intervals can provide a more informative analysis that makes it possible to view how the risk of AEs changes over the course of a clinical trial. For example, the risk of certain events may reduce as patients develop tolerability to the study medications. Alternatively, greater exposure to drug may result in an increased likelihood of certain events. Zink et al illustrate how multiple volcano plots or animation can be used to communicate the instantaneous risk within time intervals.[77] Similar presentations can be used to present analyses of cumulative risk over time. For example, guidance suggests presentations of cumulative AE rates for oncology studies at 3, 6, and 12 months, with the addition of other time points depending on the underlying nature of the disease and the duration of the trial.[16] Presentations of instantaneous risk by time interval is one way to account for and summarize the recurrence of events during the course of a clinical trial, though more formal analyses to assess the average number of events experienced over time are available.[86-91] Koch et al present a large-sample method to summarize the total number of events experienced accounting for the correlation between event frequency and patient exposure.[80]

Zhou et al raise an important point that extremely rare events should be analyzed using exact methods.[72] Xu and Kalbfleisch consider the use of propensity scores to account for differences in patient characteristics between treatments in studies of small size; propensity scores can also be applied to the comparisons of treatments within subgroups.[92,93] Given the rarity of many individual events, an alternative strategy to identify safety signals is to analyze groups of terms that describe a particular medical condition using SMQs.[8,15] Alternatively, Bayesian modeling can borrow strength across related events through their hierarchical relationships.[94-97] Other safety analyses consider the co-occurrence of events.[98,99] Site-level comparisons of individual and overall event rates can identify elevated patient risk at individual study sites.[100,101] Coupled with basket analyses, between-site comparisons of AEs can potentially uncover events that go unreported.[102]

Subgroups

Subgroups are frequently considered for the analysis of safety and efficacy endpoints, with 70% of clinical trials reporting at least some results within subgroups.[103] Subgroup analyses are beneficial in that they provide clinicians with information on the potential for differential treatment response within important demographic, genetic, disease, environmental, behavioral, or regional characteristics.[71,104] From a regulatory perspective, such analyses are important to show that the estimated overall effect is broadly applicable to patients and to assess risk-benefit across the proposed indication, particularly when the study population is heterogeneous.[105] Further, examining results within subgroups allows the study team to assess the consistency and robustness of results obtained for the entire study population, as well as to generate hypotheses for future research.[106] In trials of oncology, for example, subgroup analyses are important to identify patients at increased risk for severe toxicity of the prescribed treatments. Subgroup analyses would likely be considered for important tier I events.

When reporting results within subgroups, transparency is the key for appropriate interpretation of results. Details on subgroup size and the number of subgroups assessed (not just reported), whether subgroups were determined pre or post hoc, multiplicity adjustments were applied, stratified randomization was used, or heterogeneity was assessed through interaction tests should be clearly described.[107,108] For multiplicity, details as to whether adjusted or unadjusted *P* values are presented or simultaneous or unadjusted confidence or credible intervals should be clearly described. However, regulatory guidance appears to prefer presenting unadjusted *P* values and intervals for subgroup analyses as they are "investigations [that] serve as an indicator for further exploration."[105] Even though power tends to be low for tests of interaction, many authors suggest that heterogeneity of treatment effects should always be evaluated, and regulatory guidance encourages reporting estimates and confidence intervals for these interaction tests.[103,105,107] Further, the literature highlights that the presence and the size of interaction depends on the choice of the measure of divergence between the treatment groups.[71,105]

The measures used to determine heterogeneity should be prespecified and clearly documented. Though Figure 4 summarizes cardiovascular death or hospitalization in patients with T2DM, identical displays can be generated for important tier I events.[69] Based on recommendations from the CHMP, interaction tests are summarized using a forest plot in the right panel and are based on unadjusted 95% confidence intervals for the difference in treatment effects between the 2 subgroup levels (level 1 minus level 2, eg, metformin effect minus sulfonylurea effect).[105] Confidence intervals in the right panel that cover 0 suggest that there is no difference among treatment effects between the subgroup levels. However, this panel may help communicate (based on the width of the intervals) that there was little power available to identify a difference in the first place. It is important to note that Figure 4 presents overlapping subgroups. In other words, the same set of patients is presented, partitioned into varying subgroup levels. Alternatively, recent data-driven methodologies can be used to identify subgroups using combinations of individual factors to characterize sets of patients with differential response to treatment.[93,109-116] For safety outcomes, these methodologies can identify groups of patients for whom the new therapy may be inappropriate. See Alosh et al for a recent overview of statistical considerations for subgroups in clinical trials.[117]

Meta-analysis

While FDR can limit false positives without overly sacrificing power, the rarity of many safety endpoints will require a meta-analysis of multiple studies for sufficient power to generate meaningful inference for the safety population, as well as more precise estimates of the treatment response within various subgroups.[2,71,80,97,118] Meta-analyses should be preplanned and assess the heterogeneity and poolability of the included clinical trials using statistical methodologies, not simply reflect a naïve grouping of patients from multiple studies, since this ignores the fact that data comes from different studies. As an additional benefit, the availability of multiple trials allows the analyst to assess the consistency of response within a particular subgroup from one study to the next (ie, replication). As Li and coauthors point out, it is possible to observe negative results (even significantly so) within at least 1 subgroup when the result is known to be homogeneous among all subgroups.[119] Chuang-Stein et al provide details and recommendations for fixed and random-effects models for meta-analyses of safety endpoints.[71] Finally, it is important to note that any meta-analysis methods utilized should summarize the results of all appropriate studies to avoid biased conclusions—this includes trials where treatment arms may experience no events. For example, Marchenko et al discuss the recent Nissen and Wolski meta-analysis for T2DM.[6,120]

In 2016 the CIOMS Working Group X published a report on Evidence Synthesis and Meta-Analysis for Drug Safety.[121] The

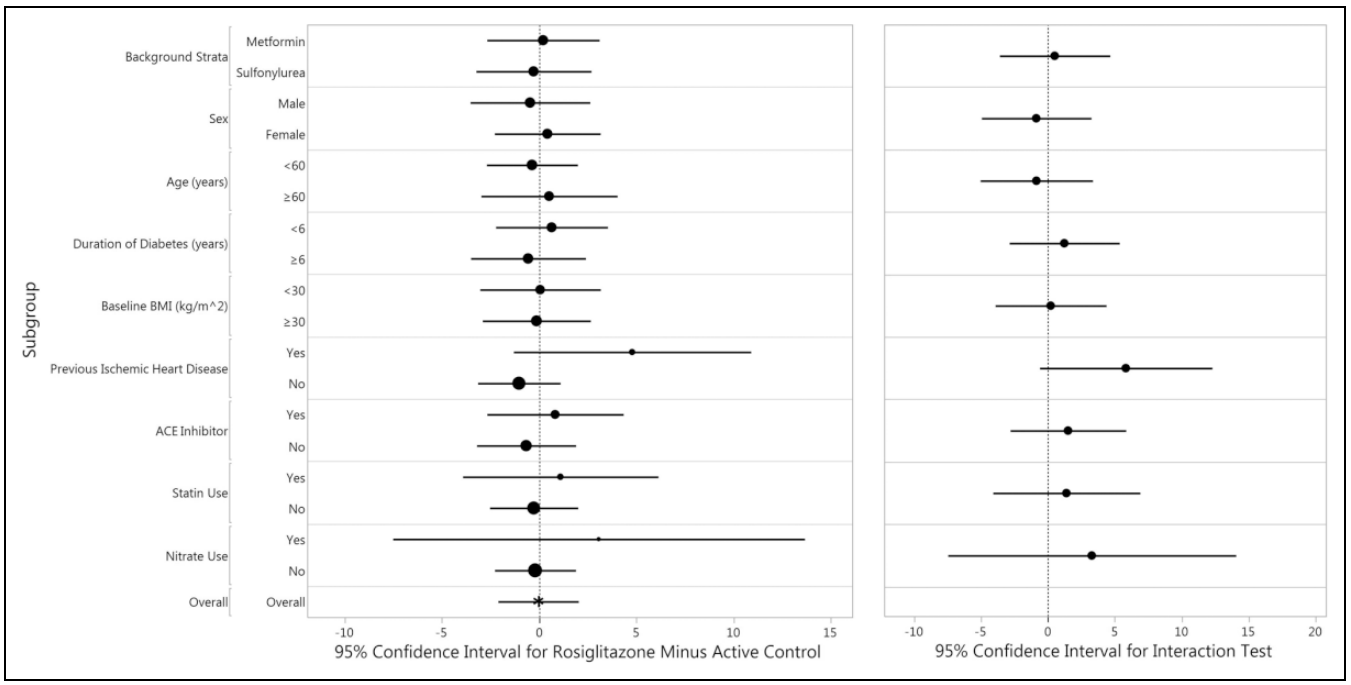

Figure 4. Subgroup analysis of cardiovascular death or hospitalization in patients with type 2 diabetes mellitus. Unadjusted 95% confidence intervals are based on the risk difference of cardiovascular death or hospitalization for rosiglitazone minus active control using a normal approximation. Interaction tests are based on unadjusted 95% confidence intervals for the difference in treatment effects between the 2 subgroup levels (level 1 minus level 2). Bubble area in the left panel is proportional to the total number of patients within each subgroup level. Data are from Home et al.[69]

goal of the CIOMS X report is to provide principles on appropriate application of meta-analysis in assessing safety of pharmaceutical products to inform regulatory decision making. Combining evidence on adverse events, where these were not the focus of the original studies, is more challenging than combining evidence on prespecified benefits. This focus on AEs represents the main contribution of the current CIOMS X report.

Advantages and Limitations of Clinical Trials for Safety Evaluation

Clinical trials are often considered the pinnacle of evidence-based medical research, and the data from safety outcomes are of high quality, ensured by prospective and uniform data collection, fastidious review and follow-up, and diligent querying and cleaning. Further, centralized labs and event adjudication are often adopted in clinical trials, which further improve the consistency of data collection and quality. Standardized medical dictionaries, such as MedDRA and WHO-DD, allow for the consistency of reporting of numerous safety outcomes.[2,14,29] Safety data collected in clinical trials are also rich and multifaceted. For example, it is possible to write detailed narratives of severe AEs that summarize the details surrounding these events to enable understanding of the circumstances that may have led to the occurrence and its subsequent management and outcome. Often these narratives include details on medical history, concomitant medications taken at the time of the event or prescribed as a result of the event, measurements of important chemical and hematology analytes or other laboratory parameters, details on hospitalizations, and whether or not the patient ultimately recovered. Direct interactions with investigators allows for the collection of data necessary to accurately describe a patient's safety profile. On the other hand, data quality, lack of important detail, and consistency of collection are several hurdles to overcome in the analysis of pharmacovigilance (PV) databases.

A concurrent control group and randomization are frequently used in clinical trials to avoid confounding, and blinding is widely used to effectively reduce bias. With patients randomized to a concurrent control group and patients, investigators and the trial sponsor blinded to individual treatment assignments, it is much more straightforward to estimate the treatment effect in a clinical trial as compared to an observational study, where it is often a nontrivial task to delineate the treatment effect from other confounding factors. Randomization has the ability to provide some balance for covariates that are unobserved or unknown to affect response. Further, unlike PV databases, clinical trials are cohort studies where the denominators are known. In other words, in PV databases there are data available for AEs for patients taking specific drugs, but not on the number of patients who do not experience events when taking certain medications. It may be possible to estimate denominators from insurance databases or sales data, though

information on demographic and disease characteristics would need to be available for useful analysis.[122] In addition, PV databases often suffer from issues like duplicated reports, underreporting of AEs, overreporting of events for new products or after public media exposure of an AE or product, and inconsistent reporting patterns.[123-125] Data in clinical trials allow for more straightforward and reliable comparison of treatments and estimation of incidence and prevalence.

While there are many benefits to the quality of safety data from clinical trials, there are serious limitations in terms of safety monitoring and assessment. It may not be feasible to power a study for safety endpoints of low frequency.[22] Further, the cost may be prohibitive and the sample size required may be too large to feasibly recruit all the subjects needed for the study.[126] Even in the SCS when data from pivotal studies are integrated for safety assessment, the sample size may not be big enough for detecting rare events and/or moderate safety shifts.[6] We advocate the importance of meta-analysis of clinical trials for safety endpoints to address the issue associated with small sample size of individual studies.[2,118,121,127] However, the use of simple pooling instead of meta-analysis in SCS is still common. In the guidance document on evaluating CV risk in therapies for T2DM, FDA laid out detailed recommendations on how to design and analyze data from multiple trials. Meta-analysis and a 2-stage assessment approach were proposed.[49]

During drug development, although toxicity studies and animal models give us insights of the safety profile of a drug, the actual clinical impact on patients may not be known before a large number of patients are exposed to the drug. For example, several drugs were pulled off from the market due to DILI. However, "the drugs that have caused severe DILI in humans have not shown clear hepatotoxicity in animals, generally have not shown dose-related toxicity, and, as noted, generally have caused low rates of severe injury in humans (1 in 5000 to 10,000 or less)."[1] Further, while the disease under investigation, the mechanism of action of study therapies, or the above toxicity studies or animal models may suggest safety issues likely to occur during the course of the trial, unplanned safety issues may emerge making it difficult to prespecify appropriate analyses in advance. Protocols and analyses have to have sufficient flexibility to address unplanned events, and the study team requires the appropriate discipline to update the PSAP as new information is learned to guide the design of future trials.

Another drawback of using premarket clinical trials for safety assessment is that subjects enrolled in clinical trials may not be representative of the general patient population. Clinical trials are designed to control variability and to ensure the quality of the generated data; therefore, the patients recruited to participate are those who meet a long list of study eligibility criteria. Concomitant therapies and confounding diseases are often listed in the inclusion/exclusion criteria. In addition, clinical trials may also require patients to be able to complete the clinical visits. Thus, clinical trials designed for efficacy enroll patients that tend to be healthier and more uniform than the general patient population. However, in order to conduct CVOTs efficiently for patients with T2DM, the enrolled patients are often sicker with a higher risk of CV events than the general population.[6] In either scenario, generalizing study findings to the larger population of patients is not possible without additional assumptions. Data from real-world sources and PV databases often reflect greater diversity and are more representative of patient experience.[4,18,128,129]

While clinical trials provide high-quality data and an initial assessment of the safety profile of a new therapy, they cannot fully characterize the safety profile on their own. Safety signals are often quite small due to the above-mentioned reasons. Thus, postmarket observational methods and real-world data sources play a critical role in further improving the safety profile of a drug.[4,18,128-130]

Conclusions

In this manuscript, the ASA Biopharmaceutical Section Safety Working Group shared its recommendations for the statistical and graphical methodologies necessary to appropriately monitor, analyze, report, and interpret safety outcomes and discussed the advantages and disadvantages of safety data obtained from clinical trials compared to other sources. As a brief summary, it is important to proactively plan for a comprehensive safety evaluation at the start of any development program, distinguishing between anticipated and unanticipated events, considering the effects of patient exposure, utilizing appropriate multiplicity adjustment and proper meta-analysis across multiple trials, and examining consistency of findings across subgroups and trials for replication of effects. Further, given the number of potential sensitivity analyses required to gain a clear picture of patient safety, the effective use of data visualization is an important consideration to efficiently summarize and review data on an ongoing basis. We further highlighted the importance of individual ethics, reminding sponsors that evaluations that limit their analyses of safety outcomes to identifying population shifts between treatments is insufficient. We encourage sponsors to regularly screen for the IMEs suggested by the EudraVigilance Expert Working Group; similar guidance for the screening of serious and unexpected suspected adverse reactions is suggested in the Federal Register and recent draft FDA guidance on safety assessment.[24,25,46,131]

Proactive planning is further encouraged in the draft FDA guidance on safety screening, which recommends that sponsors develop the appropriate procedures and infrastructure for periodic review of completed and ongoing clinical development activities for suspected adverse reactions to study therapies.[46,132] This review includes identifying events that occur more frequently in the treatment group compared to concurrent or historical controls, as well as identifying any clinically meaningful increase in rates of these events beyond what is expected. The FDA acknowledges sponsor concerns over trial integrity and the potential review of unblinded information. Therefore, to accomplish the task outlined in the guidance, the FDA suggests the formation of a SAC, a group comprised of

individuals who are independent of study teams and can review unblinded safety data to determine if reporting and further intervention is required.

All of the above discussions focus on summarizing and responding to safety outcomes as they occur. Recently, the FDA published a document outlining the initiatives and innovations to improve drug safety throughout the development life cycle.[133] For clinical trials, these efforts include identifying and validating potential biomarkers that increase the likelihood of drug-induced toxicities. For example, the Predictive Safety Testing Consortium is actively engaged in discussions with the FDA and EMA to qualify biomarkers for liver, muscle, cardiac, kidney, testicular, and pancreatic injury. In another example, the Division of Applied Regulatory Science (DARS) identified a specific protein mediator of Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS/TEN), which allowed them to identify a gene that, when disrupted, contributes to SJS/TEN. DARS is currently evaluating tools that can be used to predict the likelihood of patients experiencing drug-related AEs, including those events not previously observed in clinical trials from the compound under investigation, to help guide regulatory decision making and labeling. Similar to the initiatives and innovations document of the FDA, the ASA Biopharmaceutical Section Safety Working Group will author additional manuscripts that explore other individual sources of safety data including registries, electronic health records, and pharmacovigilance databases to describe approaches to more effectively leverage information within and between these sources.[128,129]

Readers interested in greater detail on the analysis and reporting of safety outcomes in clinical trials can explore texts by Jiang and Xia[4] or Gould,[5] or revisit Gilbert.[3] For greater therapeutic focus, readers can review a recent examination of safety specific to clinical trials in oncology.[51] Those interested in graphical presentations of safety data have numerous sources to review.[65,78,134-139]

Acknowledgments

The authors would like to thank Rima Izem, Estelle Russek-Cohen, and the reviewers for constructive comments and suggestions that improved the content of this manuscript.

Declaration of Conflicting Interests

Haijun Ma is an employee of Amgen Inc; Dr. Ma reports being employed by Amgen Inc.

Funding

No financial support of the research, authorship, and/or publication of this article was declared.

References

1. US Food and Drug Administration. Guidance for industry: Drug-induced liver injury: Premarketing clinical evaluation. https://www.fda.gov/downloads/Drugs/ . . . /guidances/UCM174090.pdf. Published 2009.
2. Crowe BJ, Xia HA, Berlin JA, et al. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. *Clin Trial*. 2009;6:430-440.
3. Gilbert GS, ed. *Drug Safety Assessment in Clinical Trials*. New York, NY: Marcel Dekker; 1993.
4. Jiang Q, Xia HA, eds. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. Boca Raton, FL: CRC Press; 2014.
5. Gould AL, ed. *Statistical Methods for Evaluating Safety in Medical Product Development*. Chichester, UK: John Wiley & Sons Ltd; 2015.
6. Marchenko O, Jiang Q, Chakravarty A, et al. Evaluation and review of strategies to assess cardiovascular risk in clinical trials in patients with type 2 diabetes mellitus. *Stat Biopharma Res*. 2015;7:253-266.
7. Marchenko O, Jiang Q, Chuang-Stein C, et al. Statistical considerations for cardiovascular outcome trials in patients with type 2 diabetes mellitus [published online February 15, 2017]. *Stat Biopharma Res*. doi/full/10.1080/19466315.2017.1280411.
8. Ma H, Ke C, Jiang Q, Snapinn S. Statistical considerations on the evaluation of imbalances of adverse events in randomized clinical trials. *Therapeutic Innovation & Regulatory Science*. 2015;49:957-965.
9. Kaizar EE, Greenhouse JB, Seltman H, Kelleher K. Do antidepressants cause suicidality in children? A Bayesian meta-analysis. *Clin Trial*. 2006;3:73-98.
10. Pinder MC, Duan Z, Goodwin JS, Hortobagyi GN, Giordano SH. Congestive heart failure in older women treated with adjuvant anthracycline chemotherapy for breast cancer. *J Clin Oncol*. 2007;25:3808-3815.
11. Suter TM, Procter M, van Veldhuisen DJ, et al. Trastuzumab-associated cardiac adverse effects in Herceptin adjuvant trial. *J Clin Oncol*. 2007;25:3859-3865.
12. US Food and Drug Administration. Pediatric study plans: Content of and process for submitting initial pediatric study plans and amended initial pediatric study plans (draft). https://goo.gl/n5zP2C. Published 2016.
13. International Conference on Harmonization. E9 (R1): Addendum to statistical principles for clinical trials on choosing appropriate estimands and defining sensitivity analyses in clinical trials. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9__R1__Final_Concept_Paper_October_23_2014.pdf. Published 2014.
14. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf*. 1999;20:109-117.
15. Mozzicato P. Standardised MedDRA queries: their role in signal detection. *Drug Saf*. 2007;30:617-619.
16. European Medicines Agency. Draft guideline on the evaluation of anticancer medicinal products in man. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2016/03/WC500203320.pdf. Published 2016.
17. Biogen Idec. Tysabri (natalizumab) injection, for intravenous use [package insert]. https://www.tysabri.com/content/dam/commer

cial/multiple-sclerosis/tysabri/pat/en_us/pdfs/tysabri_prescribing_information.pdf. Published May 2016.
18. Jiang Q, He W, eds. *Benefit-Risk Assessment Methods in Medical Product Development*. Boca Raton, FL: CRC Press; 2016.
19. Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharm Stat*. 2016;15:292-296.
20. Schnell PM, Ball G. A Bayesian exposure-time method for clinical trial safety monitoring with blinded data. *Therapeutic Innovation & Regulatory Science*. 2016;50:833-838.
21. Duke SP, Kleoudis C, Polinkovsky M, et al. Quantitative methods for safety monitoring of rare serious adverse events. *Pharm Med*. 2017;2:113-118.
22. International Conference on Harmonization. E1: The extent of population exposure to assess clinical safety for drugs intended for long-term treatment of non-life threatening conditions. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E1/Step4/E1_Guideline.pdf. Published 1994.
23. Genentech. Herceptin (trastuzumab), intravenous infusion [package insert]. https://www.gene.com/download/pdf/herceptin_prescribing.pdf. Published March 2016.
24. EudraVigilance Expert Working Group. Inclusion/exclusion criteria for the "Important Medical Events" list. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2016/08/WC500212100.pdf. Published August 2016.
25. EudraVigilance Expert Working Group. Important Medical Event Terms List (MedDRA version 19.1). http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp? webContentId=WC500208836. Published September 2016.
26. Hamasaki T, Asakura K, Evans SR, Ochiai T. Group-sequential clinical trials with multiple co-objectives. SpringerBriefs in Statistics (2016).
27. International Conference on Harmonization. M3 (R2): Guidance on nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multidisciplinary/M3_R2/Step4/M3_R2__Guideline.pdf. Published 2009.
28. World Health Organization (WHO). *The Importance of Pharmacovigilance - Safety Monitoring of Medicinal Products*. London: World Health Organization, 2002. http://apps.who.int/medicinedocs/pdf/s4893e/s4893e.pdf.
29. Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. *Drug Inform J*. 2008;42:409-419.
30. CDISC Submission Data Standards Team and CDISC SDTM Governance Committee. (2016). Study Data Tabulation Model, Version 1.5. Round Rock, TX: Clinical Data Interchange Standards Consortium. https://www.cdisc.org/system/files/members/standard/foundational/sdtm/SDTM%20v1.5.pdf.
31. CDISC Analysis Data Model Team. (2009). Analysis Data Model (ADaM), Version 2.1. Round Rock, TX: Clinical Data Interchange Standards Consortium. https://www.cdisc.org/system/files/members/standard/foundational/adam/analysis_data_model_v2.1.pdf.
32. Zink RC, Mann G. On the importance of a single data standard. *Drug Inform J*. 46 (2012): 362-367.
33. International Conference of Harmonisation. Guideline E9: Statistical principles for clinical trials. (1998). http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.
34. Ellenberg J. (1996). Intent-to-treat analysis versus as-treated analysis. *Drug Inform J*. 30: 535-544.
35. Senn S. (2007). *Statistical Issues in Drug Development, Second Edition*. Chichester, England: John Wiley & Sons.
36. United States Food & Drug Administration. (2016). Noninferiority clinical trials to establish effectiveness. https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf.
37. Lewis JA, Machin D. Intention to treat—who should use ITT? *Br J Cancer* 1993;68:647-650.
38. Wiens BL, Zhao W. The role of intention to treat in analysis of noninferiority studies. *Clin Trial*. 2007;4:286-291.
39. Sanchez MM, Chen X. Choosing the analysis population in noninferiority studies: per protocol or intent-to-treat. 2006;25:1169-1181.
40. Ioannidis JPA, Evans SJQ, Gøtzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 141 (2004): 781-788.
41. Council for International Organizations of Medical Sciences (CIOMS) Working Group VI. *Management of Safety Information from Clinical Trials*. Geneva, Switzerland. https://cioms.ch/shop/product/management-of-safety-information-from-clinical-trials-report-of-cioms-working-group-vi/. Published 2005.
42. Ellenberg SS, Fleming TR, DeMets DL. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. Chichester, West Sussex, UK: John Wiley & Sons; 2002.
43. Dixon D, Freedman R, Herson J, et al. Guidelines for data and safety monitoring for clinical trials not requiring traditional data monitoring committees. *Clin Trial*. 2006;3:314-319.
44. US Food and Drug Administration. Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees. https://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf. Published 2006.
45. US Food and Drug Administration. Guidance for industry and investigators: Safety reporting requirements for INDs and BA/BE studies. https://www.fda.gov/downloads/Drugs/Guidances/UCM227351.pdf. Published 2012.
46. US Food and Drug Administration. Guidance for industry: safety assessment for IND safety reporting (draft). https://www.fda.gov/downloads/drugs/guidances/ucm477584.pdf. Published 2015.
47. US Food and Drug Administration. Electronic Code of Federal Regulations. https://www.ecfr.gov. Published 2017.
48. Fries M, Kracht K, Li J, et al. Safety monitoring methodology in the premarketing setting. *JSM Proc*. 2016:2247-2269. https://ww2.amstat.org/MembersOnly/proceedings/2016/data/assets/pdf/389675.pdf.
49. US Food and Drug Administration. Guidance for industry: diabetes mellitus—evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. https://www.fda.gov/

downloads/Drugs/. . . /Guidances/ucm071627.pdf. Published 2008.
50. Demets DL, Lan KKG. Interim analysis: the alpha spending function approach. *Stat Med*. 1994;13:1341-1352.
51. Ivanova A, Marchenko O, Jiang Q, Zink RC. Safety monitoring and analysis in oncology trials. In:Roychoudhury S, Lahiri S, eds. *Statistical Challenges in Oncology Clinical Development*. Boca Raton, FL: CRC Press; 2018: Forthcoming.
52. Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. *Biometrics*. 1995;51:656-664.
53. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*. 1995;51: 1372-1383.
54. Conaway MR, Petroni GR. Designs for phase II trials allowing for trade-off between response and toxicity. *Biometrics*. 1996;52: 1375-1386.
55. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med*. 1995;14:357-379.
56. Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in for single-arm clinical trials. *J Clin Oncol*. 1996;14:296-303.
57. Thall PF, Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat Med*. 1998;17:1563-1580.
58. Thall PF, Cheng SC. Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics*. 1999;55:746-753.
59. Chen Y, Smith B. Adaptive group sequential designs for phase II clinical trials: a Bayesian decision theoretic approach. *Stat Med*. 2009;28:3347-3362.
60. Thall P. Some geometric methods for constructing decision criteria based on two-dimensional parameters. *J Stat Plan Inference* 2008;138:516-527.
61. Geller NL, Follmann DF, Leifer ES, Carter SL. Design of early trials in peripheral blood stem cell transplantation: a hybrid frequentist-Bayesian approach. In: Geller NL, ed. *Advances in Clinical Trial Biostatistics*. New York: Marcel Dekker; 2005.
62. Song G, Ivanova A. Frequentist enrollment and stopping rules for managing toxicity requiring long follow-up in Phase II oncology trials. *J Biopharm Stat* 2015;25:1206-1214.
63. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: CRC Press; 2000.
64. Proschan MA, Lan KKG, Wittes JT. *Statistical Monitoring of Clinical Trials*. New York, NY: Springer; 2006.
65. Pharmaceutical Users Software Exchange (PhUSE) Computational Science Standard Analyses and Code Sharing Working Group, Analysis and Display White Papers Project Team. Analysis and Displays Associated with Adverse Events: Focus on Adverse Events in Phase 2-4 Clinical Trials and Integrated Summary Documents. http://www.phuse.eu/documents//working-groups/cs-whitepaper-adverseevents-v10-4442.pdf. Published 2017.
66. Senn S. Disappointing dichotomies. *Pharm Stat*. 2003;2:239-240.
67. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. *Pharm Stat*. 2009;8:50-61.
68. Haley EC, Kassell NF, Torner JC. A randomized controlled trial of high-dose intravenous nicardipine in aneurysmal subarachnoid hemorrhage. *J Neurosurg*. 1993;78:537-547.
69. Home PD, Pocock SJ, Beck-Nielsen H, et al. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. *Lancet*. 2009;373:2125-2135.
70. Kantarjian H, Shah NP, Hochhaus A, et al. Dasatinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukemia. *N Engl J Med*. 2010;362:2260-2270.
71. Chuang-Stein C, Li Y, Kawai N, Komiyama O, Kuribayashi K. Detecting safety signals in subgroups. In: Jiang Q, Xia HA, eds. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. Boca Raton, FL: CRC Press; 2014.
72. Zhou Y, Ke Chunlei, Jiang Q, Shahin S, Snapinn S. Choosing appropriate metrics to evaluate adverse events in safety evaluation. *Therapeutic Innovation & Regulatory Science*. 2015;49: 398-404.
73. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 1995;57:289-300.
74. Benjamini Y, Yekutieli D. False discovery rate-adjusted multiple confidence intervals for selected parameters. *J Am Stat Assoc*. 2005;100:71-81.
75. Mehrotra DV, Adewale AJ. Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. *Stat Med*. 2012;31: 1918-1930.
76. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat Genet*. 2001;29:389-395.
77. Zink RC, Wolfinger RD, Mann G. Summarizing the incidence of adverse events using volcano plots and time windows. *Clin Trial*. 2013;10:398-406.
78. Amit O, Heiberger RM, Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharm Stat*. 2008;7: 20-35.
79. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *Br Med J*. 2001;322:1479-1480.
80. Koch GG, Schmid JE, Begun JM, Maier WC. Meta-analysis of drug safety data. In: Gilbert GS, ed. *Drug Safety Assessment in Clinical Trials*. New York, NY: Marcel Dekker; 1993.
81. Liu GF, Wang J, Liu K, Snavely DB. Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Stat Med*. 2006;25:1275-1286.
82. Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using SAS*. 3rd ed. Cary, NC: SAS Institute Inc; 2012.
83. Collett D. *Modelling Survival Data in Medical Research*. 3rd ed. Boca Raton, FL: CRC Press; 2015.
84. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. *Pharm Stat*. 2016;15:297-305.

85. Proctor T, Schumacher M. Analyzing adverse events by time-to-event models: the CLEOPATRA study. *Pharm Stat*. 2016;15:306-314.
86. Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics*. 1995;37:158-168.
87. Nelson W. Recurrent-events data analysis for repairs, disease episodes, and other applications. ASA-SIAM Series on Statistics and Applied Probability. Philadelphia, PA: SIAM; 2003.
88. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J R Stat Soc*. 2000;62:711-730.
89. Johnston G, So Y. Analysis of data from recurrent events. SAS User Group International, Statistics and Data Analysis 28 (2003): 1-12.
90. Diao L, Cook RJ, Lee KA. Statistical analysis of recurrent adverse events. In: Gould AL, ed. *Statistical Methods for Evaluating Safety in Medical Product Development*. Chichester, UK: John Wiley & Sons Ltd; 2015.
91. Hengelbrock J, Gillhaus J, Kloss S, Leverkus F. Safety data from randomized controlled trials: applying models for recurrent events. *Pharm Stat*. 2016;15:315-323.
92. Xu Z, Kalbfleisch JD. Propensity score matching in randomized clinical trials. *Biometrics*. 2010;66:813-823.
93. Zink RC, Shen L, Wolfinger RD, Showalter HDH. Assessment of methods to identify patient subgroups with enhanced treatment response in randomized clinical trials. In: Chen Z, Liu A, Qu Y, Tang L, Ting N, Tsong Y, eds. *Applied Statistics in Biomedicine and Clinical Trials Design: Selected Papers from 2013 ICSA/ISBS Joint Statistical Meetings*. Cham, Switzerland: Springer; 2015.
94. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*. 2004;60:418-426.
95. Xia HA, Ma H, Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *J Biopharm Stat* 2011;21:1006-1029.
96. DuMouchel W. Multivariate Bayesian logistic regression for analysis of clinical study safety issues. *Stat Sci*. 2012;27:319-339.
97. Odani M, Fukimbara S, Sato T. A Bayesian meta-analytic approach for safety signal detection in randomized clinical trials. *Clin Trial*. 2017;14:192-200.
98. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. *Proceedings of Knowledge Discovery and Data Mining International Conference*, San Francisco, August 26-29, 2001:67-76.
99. Goldberg-Alberts R, Page S. Multivariate analysis of adverse events. *Drug Inform J*. 2006;40:99-110.
100. TransCelerate Biopharma Inc. Position paper: risk-based monitoring methodology. http://www.transceleratebiopharmainc.com/assets/risk-based-monitoring/. Published 2013.
101. Zink RC, Dmitrienko A, Dmitrienko A. Rethinking the clinically-based thresholds of TransCelerate BioPharma for risk-based monitoring. *Therapeutic Innovation & Regulatory Science*. 2017:xxx-xxx.
102. Silverstein C, Brin S, Motwani R. Beyond market baskets: generalizing association rules to dependence rules. *Data Mining Knowledge Discov*. 1998;2:39-68.
103. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*. 2002;21:2917-2930.
104. Quan H, Mingyu L, Chen J, et al. Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Inform J*. 2010;44:617-632.
105. Committee for Medicinal Products for Human Use (CHMP). Guideline on the investigation of subgroups in confirmatory clinical trials (draft). European Medicines Agency. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf. Published 2014.
106. Cui L, Jung HMJ, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat* 2002;12:347-358.
107. Lagakos S. The challenge of subgroup analyses: reporting without distorting. *N Engl J Med*. 2006;354:1667-1669.
108. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357:2189-2194.
109. Battioui C, Shen L, Ruberg SJ. A resampling-based ensemble tree method to identify patient subgroups with enhanced treatment effect. *Proceedings to the Joint Statistical Meetings*, Montréal, Québec, Canada, August 3-8, 2013.
110. Dusseldorp E, Mechelen IV. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Stat Med*. 2014;33:219-237.
111. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30:2867-2880.
112. Loh WY. Classification and regression trees. *Data Mining Knowledge Discov*. 2011;1:14-23.
113. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med*. 2011;30:2601-2621.
114. Lipkovich I, Dmitrienko A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *J Biopharm Stat* 2014;24:130-153.
115. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JF. Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Stat Comput*. 2005;15:231-239.
116. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Machine Learning Res*. 2009;10:141-158.
117. Alosh M, Fritsch K, Huque M, et al. Statistical considerations on subgroup analysis in clinical trials. *Stat Biopharma Res* 2015;7:286-303.
118. Berlin JA, Crowe BJ, Whalen E, Xia HA, Koro CE, Kuebler J. Meta-analysis of clinical trial safety data in a drug development program: answers to frequently asked questions. *Clin Trial*. 2012;10:20-31.

119. Li Z, Chuang-Stein C, Hoseyni C. The probability of observing negative subgroup results when the treatment effect is positive and homogeneous across all subgroups. *Drug Inform J*. 2007;41: 47-56.
120. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*. 2007;356:2457-2471.
121. Council for International Organizations of Medical Sciences (CIOMS), Report of CIOMS Working Group X. Evidence Synthesis and Meta-Analysis for Drug Safety. Geneva, Switzerland: WHO Press; 2016.
122. DuMouchel W. Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system. *Am Stat*. 1999;53:177-190.
123. Committee for Medicinal Products for Human Use (CHMP). Guideline on detection and management of duplicate individual cases and Individual Case Safety Reports (ICSRs). http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2012/06/WC500129037.pdf. Published 2012.
124. Hoffman KB, Dimbil M, Erdman CB, Tatonetti NP, Overstreet BM. The Weber effect and the United States Food and Drug Administration's adverse event reporting system (FAERS): Analysis of sixty-two drugs approved from 2006 to 2010. *Drug Saf*. 2014;37:283-294.
125. Dimbil M, Chen D, Erdman CB, Dmakas A, Kyle RF. Adverse drug event reporting rates: comparing FAERS to clinical trials. Poster presented at: AMCP Annual Meeting, March 2017.
126. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ*. 2016;47:20-33.
127. US Food and Drug Administration. White Paper for Meta-Analyses Public Meeting (2013b). http://www.fda.gov/downloads/drugs/newsevents/ucm372069.pdf.
128. Marchenko O, Russek-Cohen E, Levenson M, Zink RC, Krukas-Hampel M, Jiang Q. (In this issue). Sources of safety data and statistical strategies for design and analysis: real world insights. *Therapeutic Innovation & Regulatory Science*.
129. Izem R, Sanchez-Kam M, Ma H, Zink RC, Zhao Y. (In this issue). Sources of safety data and statistical strategies for design and analysis: postmarket surveillance. *Therapeutic Innovation & Regulatory Science*.
130. US Food and Drug Administration. Structured approach to benefit-risk assessment in drug regulatory decision-making draft PDUFA V implementation plan—fiscal years 2013-2017. https://www.fda.gov/downloads/ForIndustry/UserFees/PrescriptionDrugUserFee/UCM329758.pdf. Published 2013.
131. US Food and Drug Administration. Final rule, investigational new drug safety reporting requirements for human drug and biologic products and safety reporting requirements for bioavailability and bioequivalence studies in humans. *Fed Regist*. (2010, Sept). https://www.gpo.gov/fdsys/pkg/FR-2010-09-29/pdf/2010-24296.pdf.
132. Xia HA, Crowe BJ, Schriver RC, Oster M, Hall DB. Planning and core analyses for periodic aggregate safety data reviews. *Clin Trial*. 2011;8:175-182.
133. US Food and Drug Administration. Drug safety priorities: Initiatives and innovation (2015-2016). (2015). https://www.fda.gov/downloads/Drugs/DrugSafety/UCM523486.pdf.
134. Chuang-Stein C, Le V, Chen W. Recent advancements in the analysis and presentation of safety data. *Drug Inform J*. 2001;35: 377-397.
135. Krause A, O'Connell M, eds. *A Picture Is Worth a Thousand Tables: Graphics in Life Sciences*. New York, NY: Springer; 2012.
136. Duke SP, Bancken F, Crowe B, Soukup M, Botsis T, Forshee R. Seeing is believing: good graphic design principles for medical research. *Stat Med*. 2005;34:3040-3059.
137. Matange S. *Clinical Graphs Using SAS*. Cary, NC: SAS Institute Inc; 2016.
138. Clinical Trials Safety Graphics Home Page. https://www.ctspedia.org/do/view/CTSpedia/StatGraphHome.
139. Duke SP, Jiang Q, Huang L, Banach M, Cherny M. Safety graphics. In: Jiang Q, Xia HA, eds. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. Boca Raton, FL: CRC Press; 2014.