Therapeutic Innovation
& Regulatory Science
2015, Vol. 49(6) 957-965

DOI: 10.1177/2168479015587363
tirs.sagepub.com

Statistical Considerations on the Evaluation of Imbalances of Adverse Events in Randomized Clinical Trials

Haijun Ma, PhD[1], Chunlei Ke, PhD[1], Qi Jiang, PhD[1], and Steven Snapinn, PhD[1]

Abstract

Adverse events (AEs) data compose the main body of safety data in clinical trials. Medically important imbalances of AEs in large double-blind randomized controlled trials (RCTs) are signals of potential adverse drug reactions. They will be further evaluated for causality and shape the initial label that gives users necessary information on the safe use of the drug. However, causality assessment in premarketing RCTs can be challenging. This article highlights key aspects that need attention and statistical analysis approaches that could be helpful for screening and evaluation of signals generated from imbalances of AEs in moderate or large RCTs.

Keywords

adverse event, clinical trial, imbalance, safety signal, adverse drug reaction

Introduction

Safety monitoring and assessment are critical to establish the benefit:risk profile of an investigational drug and to protect patient safety. Key elements in establishing the premarket risk profile of an investigational drug are derived on the understanding of the mechanism of action (MOA) and class effects, the findings in preclinical animal toxicology studies, the early safety and tolerability clinical studies in healthy volunteers and in patients, and safety data from moderate or large randomized controlled trials (RCTs). Postmarketing data further refine and improve characterization of the drug's safety profile.

Safety data in clinical trials in general consist of adverse events (AEs) reported by investigators, laboratory findings, electrocardiograms, and vital signs. It is worth noting that an AE does not have to be drug related, while adverse drug reactions (ADRs) are AEs for which "there are facts, evidence, or arguments to support a causal association with the drug."[1] Medically important imbalances of AEs are signals of potential ADRs and are subject to further evaluation for treatment causality.

For any specific events of interest identified on the basis of a drug's MOA or on prior preclinical and clinical data, it is recommended to prespecify the plan to collect and analyze them in the trial protocol. However, while prespecified events of interest are important in identifying ADRs, some ADRs may not be known in advance due to limited experience and knowledge of the drug and would thus appear unexpected. General screening of imbalances for detection of signals is difficult due to the large volume of potential events and heterogeneity in frequency and severity. In a large RCT, thousands of AEs could be reported. Numerical imbalances could be observed between treatment groups due to drug or due to chance. The goal is to efficiently screen the AEs, identify imbalances for further systematic evaluation, and decide whether any are attributable to the drug.

The choice of primary end point in phase 2 or 3 studies is typically driven by efficacy considerations rather than safety concerns. There are many challenging issues with the causality assessment of AE data in clinical trials (eg, lack of statistical power, insufficient follow-up, restricted population, large number of potential safety variables).[1-4] Medical judgment is critical for safety analyses. But what could be very helpful is the use of descriptive and inferential statistical methods to

[1] Global Biostatistical Science, Amgen Inc, Thousand Oaks, CA, USA

Submitted 02-Feb-2015; accepted 10-Apr-2015

Corresponding Author:
Haijun Ma, PhD, Global Biostatistical Science, One Amgen Center Drive, Amgen Inc, Mail Stop 24-2-C, Thousand Oaks, CA 91320, USA.
Email: hma@amgen.com

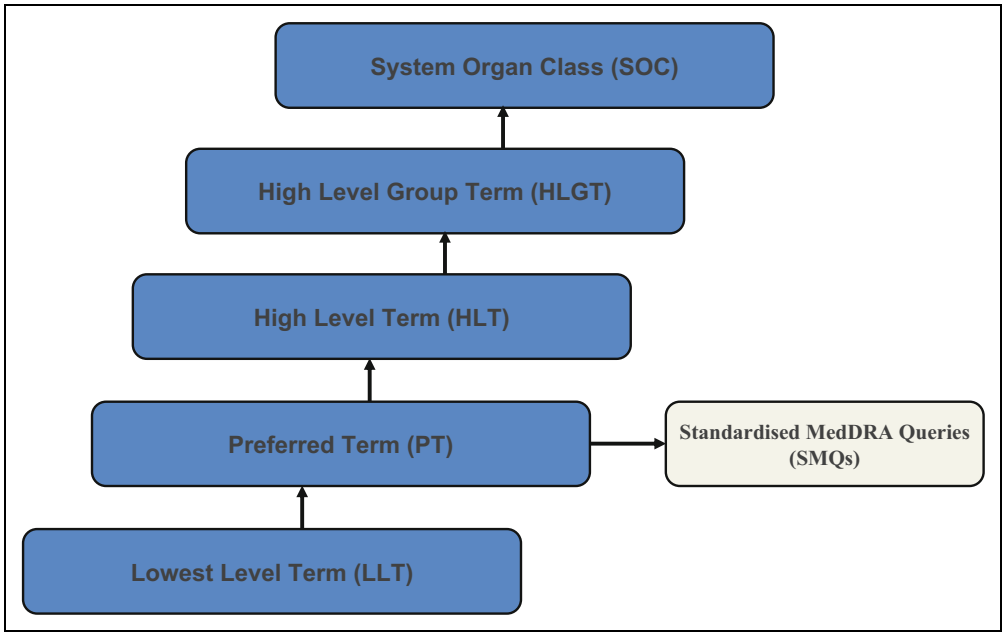

Figure 1. Medical Dictionary for Regulatory Activities (MedDRA) hierarchy.

(1) clearly describe relevant characteristics of the events and (2) decide on whether chance variation is an explanation for what is observed or whether it is more likely that some genuine drug effect has occurred.[1,3]

Although safety analysis has attracted much attention, there is still a lack of a commonly agreed systematic approach to identify and analyze safety signals from RCTs. On the basis of our experience, we found it helpful to provide guidance through a checklist of key considerations and statistical analysis methods when analyzing AE data and potential safety signals from a moderate or large RCT. In this article, we focus on identification of medically important imbalances in a moderate or large RCT and evaluation of the signals for potential ADRs in a premarketing setting. We focus on analyses for a blinded, parallel-group design that is randomized and controlled, but we also discuss considerations for open-label extensions. Methods and considerations to identify potential safety signals are provided, and we discuss statistical considerations to evaluate signals of potential ADRs, including event characteristics, patient dropout, baseline risk factors, external evidences, and so on.

Statistical Methods for Identification of Medically Important Imbalances

Filtering a large volume of AEs for medically important imbalances that warrant further causality evaluation is an important step in premarketing safety signal identification. AE screening should be based on a combination of medical judgment and statistical evidence.

Some ADRs are typically drug induced (eg, Stevens-Johnson syndrome); therefore, a small number of cases (perhaps even a single case) will be sufficient to determine causality. Some other ADRs are not commonly associated with drug exposure but are rarely seen in the population. Positive challenge-rechallenge test of individual cases may provide strong evidence of causality.

But often causality has to be determined by comparing aggregated data between treatment and control groups, especially for events that are known consequences of the underlying disease or condition under investigation or for other events that commonly occur in the study population independent of drug therapy.[5] In building the screening process of aggregated data, there are quite a few key aspects that need to be considered, including event search algorithms, statistical metrics, graphical tools, and multiplicity adjustment.

Event Search Algorithm

In clinical trials, AEs are reported using certain coding system (eg, Medical Dictionary for Regulatory Activities [MedDRA] terminology). See Figure 1 for an illustration of the MedDRA hierarchy.

Signal screening is typically performed at the preferred term (PT) level. MedDRA has 5 hierarchies with >20,000 PTs. PT tends to be granular. When multiple PTs represent the same medical concept, signals could be diluted when events are distributed across different PTs. For example, ‘‘erysipelas’’ and ‘‘cellulitis’’ are commonly seen as manifestations of the same condition and used interchangeably.[6] As a result, evaluation

at a higher level of MedDRA (eg, high-level term) is also frequently conducted.

Furthermore, standardized MedDRA queries (SMQs) have been developed to aid in the identification and retrieval of relevant safety case reports. A SMQ is a grouping of MedDRA PTs related to a defined medical condition or area of interest. Many SMQs have features that can be implemented to increase the sensitivity and specificity. SMQs have been developed for about 100 medical conditions.

Severity and seriousness are also important aspects for end point definition, where severity is often coded using Common Terminology Criteria for Adverse Events and seriousness is a regulatory definition for serious and important medical events. Adding severity and/or seriousness to end point definition allows a focus on clinically more important outcomes and/or increases specificity of the search algorithm. For example, it is common to focus on AEs of severity grade ≥ 3 for a population of advanced cancer patients.

An overly sensitive search algorithm may include too much noise, while an overly specific one may overlook true signals. Several classifications at different levels may need to be considered for screening.[3] One note worth mentioning is that "drug relatedness" of an AE reported by investigators is helpful information for causality assessment but is often influenced by the safety profile known at the time of the event. The causality assessment should be based on all AEs and SAEs irrespective of investigator-reported drug relatedness.

Descriptive Statistics and Measure of Uncertainties

AE data are often treated as binary for analysis and risk communication. This is often satisfactory for relatively short-term clinical trials.[7] The European Medicines Agency requires estimating the frequency of ADRs using crude incidence rates.[8] The treatment difference can be expressed on either a relative scale (eg, risk ratio/relative risk [RR], odds ratio [OR]) or an absolute scale (eg, risk difference [RD], attributable risk). The clinical importance of a risk depends on the potential medical consequences to patients, the prevalence of the ADR in the specific patient population being treated, and the magnitude of increased risk. For 2 ADRs with the same magnitude of RR compared to the control group, the ADR with higher prevalence will affect more patients. For example, an RR of 3 implies that the treatment causes the event in an additional 20% of patients if the reference rate is 10% but only an additional 2% of patients if it is 1%. Given the wide range of reference rates across PTs in a trial, compared to RR or, RD can more directly reflect the magnitude of patients that will be affected by the risk.[3,7,9] However, sensitivity analysis using different measures is suggested to assess robustness of conclusions.

Table 1. Sample sizes and treatment effect estimates (in percentages).

Sample Size per Arm	Pr (RR > 2.5)	Pr (RD > 7.5%)	Power of $P_1 > P_0$
2000	4	0.1	100
200	28	16	40

$P_1 = 10\%$, $P_0 = 5\%$, with true RR = 2 and RD = 5%. Pr, probability; RD, risk difference; RR, relative risk.

A commonly seen practice for identification of "common" ADR is based on RD or RR with a qualifier for risk size—that is, at least $a\%$ in the treatment group and $b\%$ greater than that in the placebo group or events occurring at an incidence of at least $a\%$ and for which the incidence is at least x times greater than the placebo incidence. However, these criteria are arbitrary and sensitive to sample size.

When sample size is small, estimates are more variable and thus less reliable. Magnitude of signals identified from small studies tends to be overestimated compared to those from big studies. Suppose that the true risks are 5% and 10% in the control and treatment groups, respectively, with a true RD of 5% and a true RR of 2. The probability of observing an estimated RR >2.5 is about 4% if the sample size is 2000 per group and 28% if 200 per group. If RD is used, the probability of observing an estimated RD >7.5% is 0.1% and 16%, respectively. However, the power to detect a difference is lower when the sample size is smaller. At a type I error level of 0.05, with a 2-sided chi-square test, the probability of detecting a higher risk in treatment group is about 100% and 40%, respectively, when sample size is 2000 and 200 per group (see Table 1).

As a consequence, for the measurement of imbalances, it is important to show uncertainty associated with the point estimates of the treatment effect. Confidence intervals could be used as a measure of statistical uncertainty due to sampling variation. For descriptive purposes, P values can also be used to show the strength of evidence against the null hypothesis of no difference between treatment and control groups. Under the Bayesian framework, posterior credible intervals and posterior probability of exceeding a threshold can be used to quantify the statistical uncertainty.

Recently, as a triage of assessment of a large amount of safety data, 7 screening criteria were proposed.[10] The criteria are a combination of inferential and descriptive statistics, ordered by importance (ie, magnitude of the RR estimate, P value, and risk size). Events meeting more important criteria are considered to show stronger evidence. The advantage of such an approach is that it provides a comprehensive search of signals and could triage the findings based on the importance of criteria met. However, the authors pointed out that medical judgment prevails and these criteria cannot be used directly to make decision.

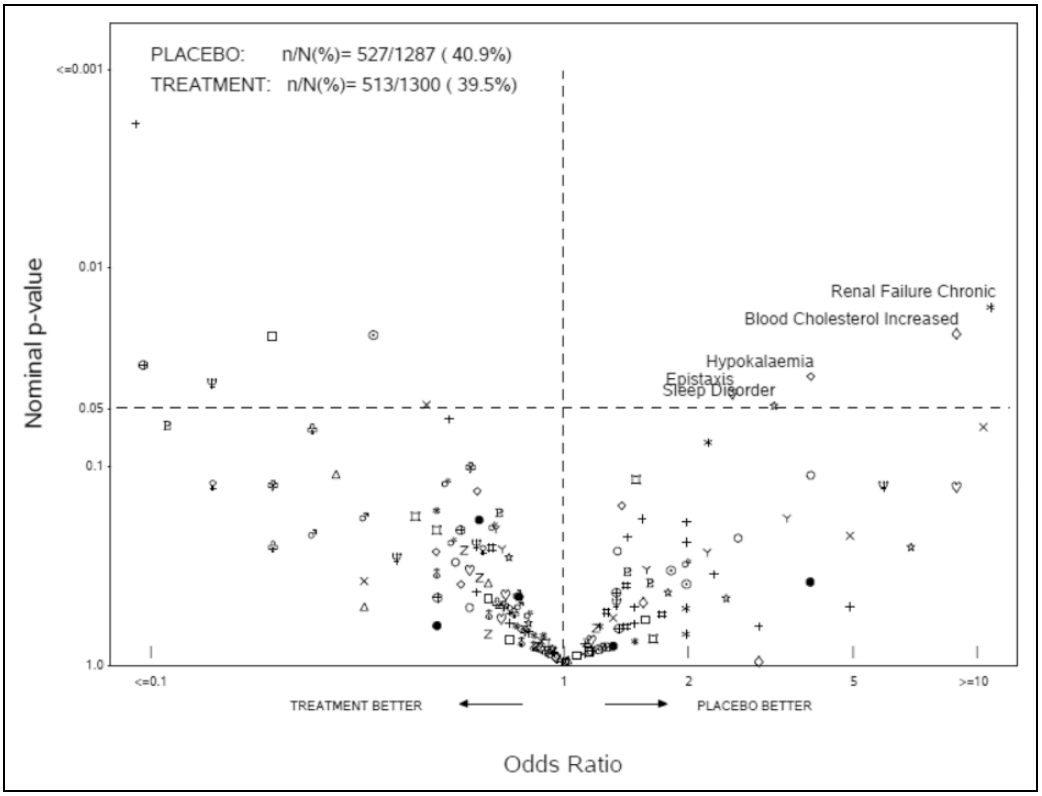

Figure 2. Volcano plot for adverse events data.

Treating an AE as binary data ignores follow-up time, event latency, or recurring events. Other statistics or methods are discussed below to address these issues.

Statistical Graphics

Statistical graphics could be very useful for signal screening, where a large amount of safety end points are examined. For example, the volcano plot in Figure 2 displays treatment effect estimates at the PT level on the *x*-axis versus nominal *P* values as a measure of statistical uncertainty on the *y*-axis.[11] Reference lines were added. The plot gives an overview of the overall balance of AEs between treatment groups. AEs falling in the upper-right corner of the plot have larger effect magnitude and smaller *P* values and are potential signals for further exploration. Other graphics have been advocated for different types of safety data.[4,12] Interactive graphics have found more usage in the pharmaceutical industry, which allow reviewers to drill down from aggregated summary to individual subjects and explore for sources of interesting patterns more easily.

Multiplicity Adjustment

A phase 3 RCT with 1000 subjects per arm could have AEs coded to >1000 unique PTs, where 5% to 15% of them may have a frequency $\geq$1%. A large number of comparisons could lead to false signals. Conventional significance levels are not suitable, and it is concerning that "any reasonable correction for multiplicity would make a finding untenable."[7] Decrease of false signals naturally causes increase of missed true signals. Magnitude of the trade-off depends on factors such as number of true signals, sample size, strength of signals, and correlation among the AE terms. Different multiplicity adjustment methods have been proposed taking advantage of MedDRA hierarchy and using Bayesian shrinkage or false discovery rate control techniques.[13-15] More research is needed in this area.

Statistical Methods for Safety Signal Evaluation

Once an AE is identified as a potential signal, it is important to perform a thorough investigation to evaluate evidence for a causal relationship with the drug. Such evaluation should include biological plausibility, nonclinical and clinical evidence, and consistency across trials of the same or similar drugs. Consistent findings across clinical trials, evidence of dose-response from population-based assessment, class effect, and/or nonclinical finding are valued as strong evidence.[7] Statistical assessment of the evidence based on randomized clinical trials is an important component of the signal evaluation, including event characteristics, dose-response relationship, risk

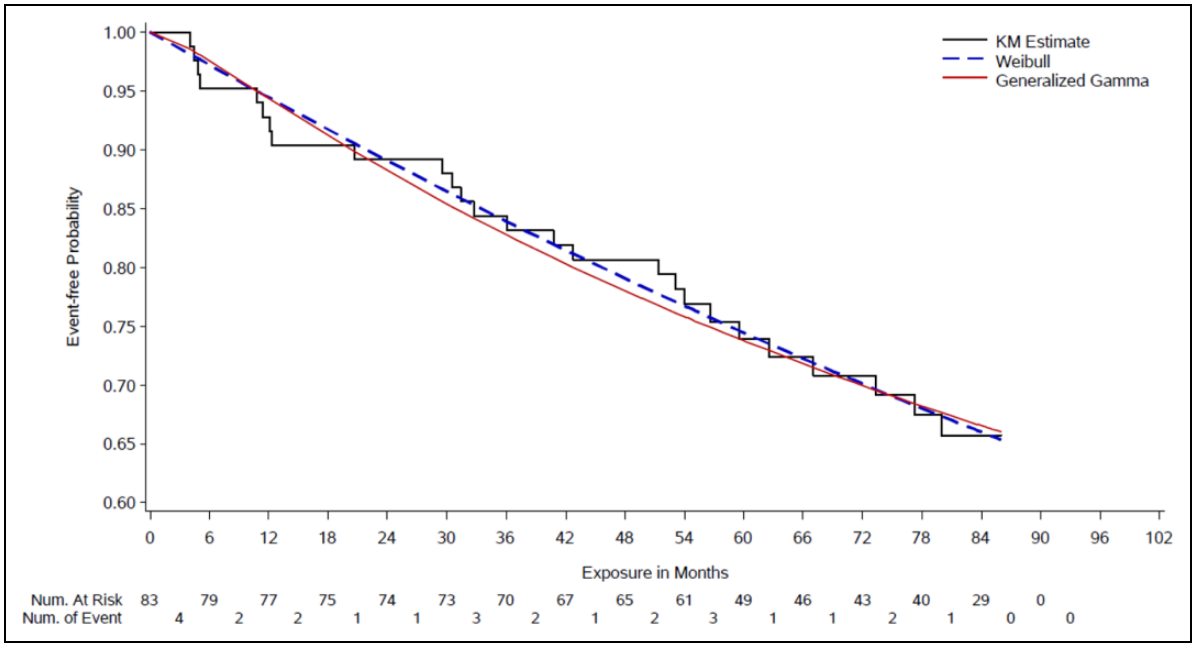

Figure 3. Kaplan-Meier (KM), Weibull, and generalized gamma fits of a long-term cohort.

factors, and other study design and conduction factors. Medical knowledge of the AEs and drug's MOA can guide the analysis and interpretation.

Adverse Event Characteristics

Evaluation of an AE's characteristics (eg, time of onset, duration, recurrence, severity, seriousness, resolution, dose-response patterns, and cause of the event) may help distinguish ADRs from AEs or further describe AEs and provide insight in risk management. For example, for injection site AEs or fever after vaccination, combining the severity or duration of AEs with the incidence could better characterize the AEs.[16,17]

Time of Event Onset

Spontaneous AEs or those related to underlying disease tend to occur randomly across the observation period, while those caused by drug exposure would display a temporal relationship. A comparison of the onset time and risk change over time between treatment and control arms is informative for causality assessment.

Analysis of the time of event onset can be based on the Kaplan-Meier (KM) estimate of the cumulative incidence or survival probability over time. A relatively constant slope of the KM curve indicates constant risk, while an abrupt increase or decrease of the slope indicates a change in risk. Estimates of hazard rate and difference of the hazard rates between treatment groups can also be used to display the change of risk over time directly, but often some smoothing techniques are necessary. An early increase of the risk may correspond to an acute ADR, whereas a chronic ADR may show only higher risk over a long exposure to drug.

Parametric distributions, such as Weibull, log-normal, and generalized gamma, are flexible and can also be used to explore the risk profile, particularly with the rare events or when there is not sufficient data. Parametric modeling allows intuitive interpretation. For example, with a Weibull fit, the shape parameter $a = 1$ indicates a constant risk; $a > 1$, an increasing risk; and $a < 1$, a decreasing risk. Figure 3 shows the KM curve and 2 parametric fits (Weibull and generalized gamma) of the time to first cardiac failure data for a long-term treatment cohort. The plot shows fairly constant risk with a corresponding Weibull shape parameter estimate of 1.018 (95% CI: 0.708, 1.463).

Some ADRs are acute (eg, hypersensitive skin reactions) and some could take a long time to develop or manifest (eg, lung cancer). Definition of the at-risk period should reflect these characteristics. For example, analysis of drug-induced allergic reactions should focus on the first few days after drug exposure. Yet, malignancy often takes time to develop; thus, latency period should be excluded from time at risk—for example, incident lung cancer identified within the first 3 months of treatment is likely not drug induced.

The assessment of bladder cancer risk associated with dapagliflozin was an example.[18] The overall cancer risk was balanced in the dapagliflozin clinical development program; analysis by cancer type showed a numeric imbalance of bladder cancer. A total of 9 (0.3%) vs 1 (0.05%) incidental bladder cancer cases were observed in the male patients. The rate ratio was

5.08 (95% CI: 0.70, 222.6; $P = .15$), but the sample size was "not powered to distinguish the incidence."

Evaluation of the disease characteristics and MOA showed that about half the cases happened within 6 months of treatment initiation, thus likely to be preexisting. Most cases were noninvasive, while one would expect more invasive cases if the drug was a tumor promoter. This observation, with other evidences, led to the conclusion that the evidence was not convincing. Due to the potential public health impact of the event, the risk of bladder cancer associated with dapagliflozin was further followed as a potential risk in postmarketing.

Recurrent Events

Often, multiple occurrences of an AE while subjects are on drug are more likely drug related. Analysis methods for multiple/repeated events could be used. A tabulation of percentage of subjects with 0, 1, 2, ... AEs during the follow-up could be used to give a basic idea of the frequency of AE recurrence. Plotting the mean cumulative function of multiple or repeated AEs per subject up to a time point can be used to show the change of event rate over time.[19] Person-time adjusted event rate is often used to provide a single-number summary of event rate, assuming constant rate over time for incidental and recurring events and ignoring intrapersonal correlation between the events. These assumptions are not always true. More complex models avoiding dependence on these assumptions, such as the Anderson-Gill model and the PWP (Prentice-Williams-Peterson) models, could be used for further exploration.[20]

Information From Other Safety Data

AEs directly reflect patients' adverse experiences during study. But due to the limited sample size and duration of clinical trials, some ADRs may not be observed frequently enough. Other safety data (eg, laboratory measurements, vital signs, use of concomitant medications) could provide important insight into the safety profile of a drug, especially when validated biomarkers exist.

Mean trend changes of biomarkers and/or outlying observations could be very revealing. Side-by-side box plot of changes in blood pressure can be examined for increasing trend as an indication of potential cardiovascular risk. Line plots of calcium can be compared across treatment groups to aid assessment of risk of hypocalcemia. Simultaneous elevation of aminotransferase and bilirubin helps identification of potential Hy's law cases for assessment of drug induced liver injury.

Patient profile plots display different types of safety information of a patient in one plot and could be used to facilitate review of individual cases.

Table 2. Summary of study dispositions, No. (%).

	Treatment	Control
Randomized subjects	174	170
Completed randomized treatment	130 (75)	151 (89)
Withdrew during randomized treatment	44 (25)	19 (11)
Reasons for discontinuation		
Adverse event	16 (9)	0
Physician decision[a]	3 (2)	0
Withdrawal by subject[a]	21 (12)	8 (5)

[a]5 dropouts in the treatment arm due to inadequate efficacy and 1 due to adverse event.

Dropout Rate and Pattern

In clinical trials, some patients drop out from treatment and/or from study follow-up for various reasons (eg, lack of efficacy, AE, loss to follow-up). Dropout due to AE is considered an important safety end point, which could be a clue to unexpected but important ADRs. However, dropout from study terminates AE data collection and creates a missing-data problem.

Reasons for dropout should be investigated and compared between treatment groups using tables or plots of cumulative dropout rates by reasons. Any clinically meaningful difference between treatment groups should be investigated. Table 2 is an example of a study disposition summary table. Discontinuations were more frequent in the treatment vs control. Reasons for discontinuation were further examined. It was found that disproportionally more treated subjects discontinued due to AEs or inadequate efficacy than did subjects in the control group.

When potential ADRs are analyzed, methods such as exposure-adjusted analysis, KM, and Cox regression can be used, assuming that dropouts are independent of end points. In the case of dependent censoring where the dropout is related to the risk of an AE, more complicated statistical analyses, such as inverse probability of censoring weighting, can be used to account for the effect of dropout.

Baseline Risk Factors

Risk factors are patient characteristics or factors associated with an increased probability of developing a condition or disease. Randomization achieves approximate balance among treatment groups in terms of known and unknown risk factors. However, a chance imbalance of risk factors is still possible. Imbalance of strong risk factors could contribute to the observed AE imbalance and cause confounding effect in risk assessment. Confounding can be a bigger issue when randomization is not maintained—for example, in an open-label extension phase of a RCT.

Whereas these analyses are post hoc, to reduce biases, it is preferred to predefine analysis methods and the set of risk

factors (eg, demographic, disease characteristics, concomitant medications, medical history, and laboratory values) based on biological and external evidence. Stratified or subgroup analysis can be conducted to examine whether treatment effect differs across the levels of risk factors, geographic regions, or centers. The nature of an interaction should be investigated. For example, it may be possible that the safety issue is limited to a specific subgroup. However, due to the post hoc nature, findings should be interpreted in the context of biological plausibility and consistency with external evidence and considered as hypothesis generating. Graphic approaches can also be used to visually examine treatment effect estimates across subgroups. Statistical tests for a quantitative and/or qualitative interaction can also be used, although it is recognized that it often has a low power. Multiple regression models can be used to adjust for the imbalance in risk factors.

External Evidence

A safety signal will likely be real if a similar issue exists in other studies or for products in the same class (class effect). However, the imbalance is more likely to be a chance finding if it is not repeated in other similar programs or studies. Consistent patterns are considered a strong evidence for causality. Nonclinical findings can also provide strong evidence. Consideration of the safety findings in individual studies must be integrated with the overall safety experience, taking into consideration all data and analyses pertinent to the issues.[7]

In the example of the potential risk of bladder cancer associated with dapagliflozin, extensive analyses were conducted. Cases of treated subjects were compared to the Surveillance, Epidemiology, and End Results database and found to exceed the expectance for diabetic population. But there was a lack of carcinogenic evidence in preclinical data and the clinical program for canagliflozin, a drug from the same class. All this information was considered and played an important role in the assessment of the safety issue.

Descriptive and graphical summaries (eg, side-by-side bar plot, forest plot) are helpful to display evidence from other sources. Meta-analysis methods, including Bayesian methods, can be used to synthesize information from different studies and is particularly useful for safety evaluation in the regulation of pharmaceutical products.

Studies With Open-Label Extension Phase

Sometimes the follow-up of an RCT is extended in an open-label phase that allows switching treatments or puts all subjects on the new treatment. Drop-in is a similar issue where control subjects are allowed to roll over to receive treatment after disease progression. Both randomization and blinding are lost in this phase. Potential confounding in such a design is discussed by Rothman,[21] where a stratified analysis by period was suggested. Techniques described in the ''Baseline Risk Factors'' section that deal with imbalanced risk factors could also be used. Knowing the treatment received by individuals in the extension phase may cause bias in assessment/reporting of outcomes of interest (eg, ascertainment bias) and provision of supplemental care or concomitant medications. The bias is often much less if the end points can be measured objectively. The direction or magnitude of the bias could be estimated by comparison to other studies or the original treatment phase, with the caveat that the comparisons may suffer from heterogeneity between studies or confounding of time-varying patient characteristics (eg, age).

When all subjects receive the new treatment in the extension phase, there is no control arm. The safety analysis should be primarily based on the randomized treatment phase. But the long-term impact of treatment can be evaluated by examining the change of risk using all data from the original treatment phase throughout the extension phase. For instance, Figure 3 shows the risk of cardiac failure with chronic use of a treatment, using data from the initial randomization through extension.

Comparing subjects in the open-label phase based on their original double-blinded phase treatment assignment is seen in practice. The comparison is actually between patients exposed to treatment for a long time and those newly exposed, and it equally suffers from the confounding issue described previously.

Additional Analyses

In addition to the previously discussed aspects, there could be other sources of information—for example, review of product quality testing by lot number analysis (CMC [chemistry, manufacturing, and controls] analysis). When local laboratories are used, it helps to check whether assays from the same manufacturer are used and whether the reference ranges and test characteristics are similar.

Concomitant medications taken and/or medical procedures performed during study may have an impact on outcome. On-study laboratory values may provide information to better understand etiology of the event. However, these postrandomization time-dependent variables may be affected by treatment. Adjusting for them in a standard regression model may obscure the true treatment effect. Similarly, subgroups identified using on-study information could be misleading for treatment effect assessment. Caution should be exercised to understand the role played by them in the causal pathway and the goal of the analysis.[22,23]

Summary and/or graphic display of AE by on-study information (collected before event) and by treatment group could

be used to understand the data. Advanced statistical modeling approaches taking into account on-study information can be used to fully utilize available information to further understand the safety data—for example, time-varying covariate models, generalized estimating equations, and joint modeling of longitudinal data and time-to-event data.

Discussion and Concluding Remarks

RCTs are often considered as of top quality in the pyramid of evidence-based medical research. The strength of clinical trials for AE assessment includes prospectiveness of the data collection, the proper and balanced control group, and blinding, which reduces bias.[24] While clinical trials provide a highly regulated environment to remove confounding factors that could influence the treatment effect estimation for efficacy, causality assessment for AEs in clinical trials can still be difficult, especially for unexpected safety issues. Underdetection due to lack of power and false alarms due to multiple testing are two important statistical challenges in premarketing safety assessment, among others. Careful evaluation of clinical trial data sets should provide further insights to facilitate the evaluation. The biological plausibility and consistency with external evidence are important considerations to evaluate those signals. In this article, we review signal screening approaches and provide systematic statistical methods to evaluate factors that may contribute to the imbalance of an AE. Medical judgment (eg, case confirmation, biologic plausibility, MOA) was crucial in safety assessment, and analytic tools helped to quantify the evidences.

Statisticians play a key role in developing tools to help clinicians efficiently identify signals for further evaluation, in interpreting clinical questions, and in addressing them through appropriate statistical analyses. Clinicians and statisticians should exercise medical and analytic judgment in interpreting analysis results.

Although the recommendations focus on issues and analyses after study unblinding, many of the analyses could and should be considered during the study design stage. If applicable, preplanned analyses are strongly recommended over post hoc analyses. Building a process and tools proactively could improve transparency, increase efficiency, and sustain the objectivity of safety analysis. However, due to the limitations of clinical trials, a definitive conclusion cannot always be made using clinical trial data alone, and postmarketing data are helpful to gather more evidence to refute or confirm the signal and further characterize the safety profile of a drug.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article. All authors are employees and stock holders of Amgen Inc.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Council for International Organizations of Medical Sciences. *Management of Safety Information from Clinical Trials*. Geneva, Switzerland: Council for International Organizations of Medical Sciences; 2005.
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. *Clinical Trials*. 2009;6(5):430-440.
3. Talbot J, Aronson JK. *Stephens' Detection and Evaluation of Adverse Drug Reactions Principles and Practices*. 6th ed. New York, NY: John Wiley & Sons Ltd; 2012.
4. Jiang Q, Xia HA. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. New York, NY: Taylor & Francis; 2015.
5. Food and Drug Administration. *Guidance for Industry and Investigators, Safety Reporting Requirements for INDs and BA/BE Studies*. Silver Spring, MD: Food and Drug Administration; 2012.
6. Kilburn SA, Featherstone P, Higgins B, Brindle R. Interventions for cellulitis and erysipelas. *Cochrane Database Syst Rev*. 2010;6: CD004299.
7. Food and Drug Administration. *Attachment B: Clinical Safety Review of an NDA or BLA of the Good Review Practice. Clinical Review Template (MAPP 6010.3 Rev. 1), Dec 15, 2010*. Silver Spring, MD: Food and Drug Administration; 2010.
8. European Medicines Agency. *A Guideline on Summary of Product Characteristics (SmPC), Revision 2, Sept. 2009*. London, England: European Medicines Agency; 2009.
9. Zhou Y, Ke C, Jiang Q, Shahin S, Snapinn S. Choosing appropriate metrics to evaluate adverse events in safety evaluation. *Therapeutic Innovation & Regulatory Science*. In press.
10. Crowe BJ, Brueckner A, Beasley C, Kulkarni P. Current practices, challenges, and statistical issues with product safety labeling. *Statistics in Biopharmaceutical Research*. 2013;5(3):445-454.
11. O'Connell M, Knudsen S. Statistical graphics and reporting in drug development. Paper presented at: PhUSE 2006; October 9-11, 2006; Dublin, Ireland. Paper TS04.
12. Amit O, Heiberger RM, Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics*. 2008;7(1):20-35.
13. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*. 2004;60:418-426.
14. Xia HA, Ma H, Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *J Biopharm Stat*. 2011; 21(5):1006-1029.
15. 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.
16. Chang MN, Guess HA, Heyse JF. Reduction in burden of illness: a new measure in prevention trials. *Stat Med*. 1994;13:1807-1814.

17. Su L, Tucker R, Frey SE, et al. Measuring injection-site pain associated with vaccine administration in adults: a randomized, double-blind, placebo-controlled clinical trial. *J Epidemiol Biostat*. 2000;5:359-366.
18. Food and Drug Administration. *Transcript for the Meeting of the Endocrinologic and Metabolic Drugs Advisory Committee for Dapagliflozin NDA, December 12, 2013*. Silver Spring, MD: Food and Drug Administration; 2013.
19. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. *J Biopharm Stat*. 2009;19(5):889-899.
20. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika*. 1981;68(2): 373-379.
21. Rothman KJ. A potential bias in safety evaluation during open-label extensions of randomized clinical trials. *Pharmacoepidemiol Drug Saf*. 2004;13(5):295-298.
22. Rosenblum M, Jewell NP, van der Laan M, Shiboski S, van der Straten A, Padian N. Analysing direct effects in randomized trials with secondary interventions: an application to human immunodeficiency virus prevention trials. *Journal of the Royal Statistical Series A*. 2009;172(2):443-465.
23. Dai JY, Gilbert PB, Mâsse BR. Partially hidden Markov model for time-varying principal stratification in HIV prevention trials. *J Am Stat Assoc*., 2012;107(497):52-65.
24. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. 4th ed. New York, NY: Springer; 2010.