Therapeutic Innovation
& Regulatory Science
2014, Vol 48(1) 109–120

DOI: 10.1177/2168479013510917
tirs.sagepub.com

Statistical Evaluation of Drug Safety Data

H. Amy Xia, PhD[1], and Qi Jiang, PhD[1]

Abstract

There has been growing awareness of the importance of the statistical evaluation of drug safety data both in the premarketing and postmarketing settings. Careful and comprehensive approaches are warranted in safety evaluation. This paper offers a high-level review of some key issues and emerging statistical methodological developments. Specifically, the following topics are discussed: prospective program-level safety planning, evaluation, and reporting; the impact of adverse event grouping on statistical analysis; the applications of Bayesian methods in safety signal detection; meta-analysis for analyzing safety data; and safety graphics. Aspects related to benefit-risk assessments are also covered.

Keywords

PSAP, adverse event grouping, Bayesian methods, signal detection, meta-analysis, safety graphics, benefit-risk assessment

Introduction

There has been growing awareness about the importance of statistical evaluation of drug safety data both in the premarketing and postmarketing settings. Over the last decade, numerous regulatory guidance documents have been issued to call attention to drug safety evaluation both at the national and international level. Examples include the following:

1. ICH E1-E2F on population exposure, definitions, standards for clinical safety data reporting and transmission, pharmacovigilance, and periodic safety update reports[1]
2. Council for International Organizations of Medical Sciences (CIOMS) VI report on the management of safety information from clinical trials (2005)[2]
3. FDA guidances (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm):
   - Premarketing Risk Assessment (2005)
   - Development and Use of Risk Minimization Action Plans (2005)
   - Good Pharmacovigilance and Pharmacoepidemiology Practices (2005)
   - Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review (2005)
   - ICH E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non Antiarrhythmic Drugs (2005)
   - Diabetes Mellitus: Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes (2008)
   - Drug-Induced Liver Injury: Premarketing Clinical Evaluation (2009)
   - Determining the Extent of Safety Data Collection Needed in Late Stage Premarket and Postapproval Clinical Investigations (2012)
   - Safety Reporting Requirements for INDs and BA/BE Studies (2012)
   - Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data (2013)
   - Providing Postmarket Periodic Safety Reports in the ICH E2C(R2) Format (Periodic Benefit-Risk Evaluation Report) (2013)
4. European Commission's detailed guidance on the collection, verification, and presentation of adverse event/reaction reports arising from clinical trials on medicinal products for human use (CT-3) (2011)[3]

There is also a general acknowledgment that sponsors need to think about their safety assessment strategies for a drug development program earlier and more proactively. A systematic, consistent approach is needed for safety planning, data collection, evaluation, and reporting.[4]

[1] Global Biostatistics Science, Amgen, Thousand Oaks, CA, USA

Submitted 6-Jun-2013; accepted 8-Oct-2013

Corresponding Author:
H. Amy Xia, PhD, Global Biostatistics Science, Amgen, One Amgen Center Drive, MS 24-2-A, Thousand Oaks, CA 91320, USA.
Email: hxia@amgen.com

Safety data are multidimensional and complex, including but not limited to adverse events (AEs), laboratory test data, electrocardiogram (ECG) data, vital signs, and other relevant assessments that are important in evaluating safety data such as demographics, medical history, and others. Careful and comprehensive approaches are warranted in safety evaluation and analysis.

Current statistical approaches for analyzing safety data are often descriptive and perhaps oversimplified, and knowledge of and experience with proper methods may be inadequate. There are some statistical challenges in quantitative safety analyses[2]:

- Power: Most of the studies in a drug development program are designed based on efficacy endpoints as primary endpoints and may have limited power to detect important differences in safety endpoints. The problem is more profound, and the false-negative rate could be high when uncommon or rare AEs occur. Most individual studies, or even a combination of trials, tend to be too small.
- Multiplicity: In clinical trials, because the number of AE types can be very large (typically in the hundreds or even thousands in late-phase clinical trials), in general, it is difficult to prespecify hypotheses for safety events. The current frequentist-based approaches of flagging AEs based on unadjusted *P* values or confidence intervals (CIs) can result in an excessive number of false-positive signals. On the other hand, if we adjust for multiplicity in traditional ways (eg, the Bonferroni method), this may lead to an excessive rate of false-negative findings, whence important safety signals may be missed, which may have greater public health implications.
- Medical classification: The grouping of AEs into categories for events of interest represents a statistical challenge. If they are too narrow, it not only affects statistical power for the comparison between groups but also can fail to group events that are medically related, and consequently, a potential signal might be missed. In contrast, if the groupings are too wide, it could mask a real safety signal by introducing potential noises. Another challenge in grouping is to form a medical concept for analyzing an event of interest so that results can be reliably interpreted.
- Complexity of safety data: Safety data often include many elements. To evaluate multidimensional, interrelated complex safety information as a whole poses a statistical challenge.

A new important risk usually prompts a benefit-risk (B-R) assessment as the impact of the new signal must be reviewed and weighed in the context of the B-R profile of the drug. How to evaluate benefits and risks qualitatively and/or quantitatively is a challenging statistical and decision-making problem.

This paper offers a high-level review of some key issues and emerging statistical methodologies in drug safety evaluation. Specifically, we include the following topics: prospective program-level safety planning, evaluation, and reporting (section 2); the impact of AE grouping on statistical analysis (section 3); the applications of Bayesian methods in safety signal detection (section 4); meta-analysis for analyzing safety data (section 5); and safety graphics (section 6). In addition, we cover some aspects related to B-R assessments (section 7). Finally, we offer some concluding remarks in section 8. The challenges in drug safety assessment are multifaceted; therefore, we believe that it is valuable to summarize various aspects relative to statistical safety evaluations in a single paper. Although many topics covered in this paper primarily focus on safety evaluation in premarketing drug development, we also discuss various aspects related to postmarketing activities, for example, the aspects related to different postmarketing sources of data for signal detection. Certainly, a B-R assessment is continuous throughout the drug development's life cycle.

Program Safety Analysis Plan (PSAP)

It is critical to proactively plan for the evaluation of safety data and to ensure that safety signals are detected in a timely manner. Having a PSAP can help achieve this goal. Although having a PSAP is not required by regulatory agencies at this time, it is an important tool to help sponsors consider how to plan for what data to collect and how to analyze and interpret the safety data throughout the life cycle of drug development. This has been recommended by the Safety Planning, Evaluation and Reporting Team (SPERT) as well, under the auspices of the Pharmaceutical Research and Manufacturers of America (PhRMA), and is becoming a good practice in industry.[5] Thus, what is a PSAP? A PSAP is a living document (amended as needed through the product's life cycle) that provides a systematic way to assess prospectively defined safety outcomes as well as to identify safety signals at a program level. It has both prospective (eg, minimum critical toxicities and AEs of special interest [AESIs]) and retrospective (eg, unexpected, late-emerging safety issues) aspects. The PSAP is maintained by the multidisciplinary safety management team.

The key components of a PSAP template could include the following: background, general plan, data generation, data structure and content, methods for analysis, presentation and reporting, and problem-oriented summary for AESIs. There are many benefits of having a PSAP:

- Be proactive and plan early for safety assessment at the program level;
- Have a systematic and consistent approach for planning, analysis, and reporting of safety data in clinical trials;
- Identify potential risks earlier in the drug development process to allow data collection strategies to be modified in time to collect additional data to further understand a safety issue;
- Facilitate communications with regulatory agencies regarding key safety evaluations for products in phase 2/3 development, and reach an agreement with agencies early if needed;
- Facilitate ongoing safety assessments throughout the life cycle of drug development,[6] and permit ongoing refinement of the understanding of the B-R profile of a new product during the postapproval phase; and
- Meet the new industry standard for safety assessment.

One of the main purposes of the PSAP is to allow teams to plan early and be proactive. To fully realize the benefits of a PSAP, development of a PSAP should be initiated during phase 2's product development in preparation for the end of phase 2 portal. It is recommended that the key components of the PSAP may be discussed with the FDA and other regulatory agencies during the end of phase 2 meetings.

The development of a PSAP is a multidisciplinary collaboration. For example, the 2 key authors are from biostatistics and safety with contribution from many disciplines, including clinical and regulatory. It would be helpful to have a PSAP's development incorporated into project timelines for appropriate time and resource planning.

A question often raised is why a PSAP is needed, since much of the content of the PSAP is covered by other documents, such as the risk management plan (RMP) and the Statistical Analysis Plan for Summary of Clinical Safety (iSAP). The PSAP is related to other documents but is somewhat different:

- The PSAP complements the RMP and specifies the analysis of the safety data in more detail (similar to the statistical analysis plan complementing the protocol). The development of a PSAP often occurs concurrently with the RMP's development.
- The PSAP serves as a basis for development of an iSAP, but there are some important distinctions. The PSAP is a living document and is being maintained throughout a product's life cycle. Per the recommendation of the SPERT, the PSAP will eventually form the basis for key analyses contained in the iSAP, and the iSAP should be consistent with the key components of the PSAP. That is, once the PSAP is approved, individual protocol-specific statistical analysis plans and the iSAP can reference the PSAP for key elements of safety data collection and analysis. The PSAP needs to be put together at an early stage of drug development. In contrast, the iSAP has a "cross-sectional" feature and, in general, is put together much later when phase 3 studies are being conducted. In addition, the PSAP will need to be updated through the drug's life cycle, but the iSAP's efforts will be over once the filings are completed.

A PSAP is critical for proactive safety evaluation and signal detection. Creating a standard operating procedure for a PSAP and/or a PSAP template would often help thinking and implementation.

Impact of AE Grouping on Statistical Analysis

The lack of standard definitions of AEs, coding conventions, and terminology usage and the lumping/splitting of terms without prospective plans can obscure signal detection and evaluation. This also is one of the challenges that data monitoring committees are generally facing in assessing safety during clinical trials.[7] Grouping AEs into categories can pose statistical challenges. Historically, a broad search is viewed as conservative in the sense that it minimizes the risk of missing an event of interest. However, it can lead to nondifferential misclassification and bias the relative risk estimate toward the null. As a result, a treatment effect can be diluted, and a signal can be masked.[8]

Figure 1 shows a vivid example of cardiovascular (CV) events from a trial studying celecoxib and placebo.[9,10] In the figure, CV events are classified in a hierarchy in the x-axis from narrow to broad (from left to right). The y-axis on the left represents the treatment effect as a hazard ratio (HR) and its corresponding 95% CI. The y-axis on the right indicates *P* values. Each successive event added to the hierarchy is considered less clearly in the pathophysiological pathway. From the very left, CV death is defined as a "pinnacle" event, and then myocardial infarction (MI), stroke, congestive heart failure (CHF) angina, and CV procedures are added in a successive way. Moving from left to right in the hierarchy, one adds events that are successively less likely to be related to the underlying mechanism of action or that by their nature are more likely to be misclassified in clinical trials. There are statistical implications as one moves from the left to the right in the hierarchy: HRs may decrease monotonically. As more likely unrelated events are added in, it is likely to bring the HR toward the null. The CIs decrease as more events are added. Statistical power, as manifested in the *P* values (see the dotted line in the figure), which are functions of both sample size and effect size, may first decrease as more significant results are observed, resulting from the increased number of events, but then increase. At the

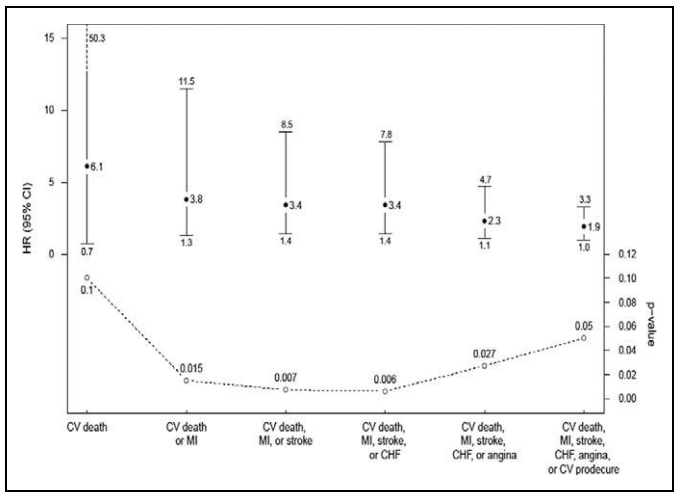

Figure 1. Hierarchical outcome classification. An example with cardiovascular events.

very right, even though the number of events is increased, the increase in noise makes the *P* values become larger than those in the middle. This example suggests that somewhere, there is the right balance for the definition of AEs. From a practical standpoint, it might be useful to investigate a variety of definitions and look for sensitivity of the conclusions against different definitions, especially in a situation where there is no existing standard definition for a particular medical event.

The Medical Dictionary for Regulatory Activities (MedDRA) consists of clinically validated international medical terminology used by regulatory authorities and the regulated biopharmaceutical industry. The terminology is used through the entire regulatory process, from premarketing to postmarketing, and for AE data entry, retrieval, evaluation, and presentation. Its subject matter comprises signs, symptoms, diseases, diagnoses, therapeutic indications, results of investigations, procedures, and medical/social/family histories.[11] Medical terms form a natural hierarchy of generality, which is useful in categorizing and referring to medical concepts. The MedDRA hierarchy consists of 5 levels from the most granular to the least: lower level term, preferred term (PT), high level term, high level group term, and system organ class (SOC). A difficulty arises in deciding whether groupings of different event terms for a patient can be formally regarded as a medical concept. Two or more PTs may stand for the same medical concept. In addition, some terms are more general than others. Inconsistencies in the classification and/or codification of clinical events are common among investigators who report AEs within the same study, among sponsors using different AE coding terminologies/dictionaries, and even among sponsors using the same dictionaries. The use of different coding dictionaries or different levels within those dictionaries can result in the aforementioned problems related to misclassification and power. One way to overcome this issue is to use standard medical definitions of AEs such as standard MedDRA queries when available. If there is no well-established definition in the medical literature, it is important to reach regulatory agreement for the use of a nonstandard definition. As mentioned in section 2, a PSAP can be used to discuss the definitions for AESIs at milestone meetings with regulatory authorities. Furthermore, the results of the analysis will need careful interpretation rather than a pure reliance on a statistical test through a close collaboration between clinicians and statisticians.

Bayesian Applications in Safety Signal Detection

Data on AEs can be classified into 3 categories: tier 1 events are those AEs with prespecified hypotheses, and tier 2 and 3 AEs are those without.[12] The distinction between tier 2 and tier 3 events is that tier 2 events are relatively common, while tier 3 events are relatively uncommon or rare. Safety data can stem from a variety of different data sources, ranging from clinical trial data, spontaneous AE-reporting databases maintained by regulatory authorities (eg, the FDA's Adverse Event Reporting System [AERS] and Vaccine Adverse Event Reporting System and the EMA's EudraVigilance), as well as observational databases including electronic health records and claims. Although different data sources have their own unique characteristics and analytical issues, how to analyze a large number of nonspecific tier 2 and 3 AEs is a common statistical challenge regardless of the data sources. The goal for data analysis is signal detection, that is, to identify certain possible risks of AEs for further investigation. In this setting, multiplicity and rare events are the most challenging problems that statisticians encounter.

Bayesian methods have many advantages in the area of safety signal detection. First, AEs usually are coded with an existing coding dictionary (eg, MedDRA for clinical trial data or ICD-9 for claims database). Rather than considering each type of AE independently, it allows for explicitly modeling AEs with the existing AE coding structure, so that strength can be borrowed within and across certain levels within the coding hierarchy. A nice feature of Bayesian hierarchical modeling is that it provides "partial correction" that accounts for multiplicity when it is crucial (ie, avoid detecting abundance of false-positive results) but does not overdo it when it is not (ie, let data determine how much borrowing would incur).[13] It could improve signal discrimination by reinforcing or tampering a signal depending on the group's behavior.[14] This is especially appealing in the rare event setting since it analyzes the entire AE dataset and modulates the extremes. For example, if many medically related AEs under the same SOC had AE counts of 3 versus 0 in the treatment and control arms, respectively, it would strengthen the signal by borrowing strength among

different AEs within the SOC in the Bayesian approach, while if only the individual cases were considered, it is possible that the signal could have been missed. Second, the Bayesian approach is attractive statistically in dealing with rare AE data because the model adaptively modulates the extremes. The inferences are based on the full posterior distributions, relaxing the need to assume normality, which is commonly assumed but may not be sensible for rare events. Third, Bayesian methods offer ease of interpretation. It is straightforward to assess the posterior probability of clinically important differences on different scales (risk difference, odds ratio, or relative risk) in order to avoid detecting medically unimportant signals. Finally, it makes efficient use of all the data. Distinction of tier 2 and tier 3 events is not necessary with the Bayesian approach, in contrast with other frequentist approaches such as the double false discovery rate method.[15]

A number of Bayesian approaches have been used in analyzing different safety data. In the clinical trial arena, a novel Bayesian hierarchical modeling approach is proposed to analyze binary outcomes in AE data.[16] The approach models AEs within the coding hierarchy so that AEs within and across, say, a SOC can borrow strength from each other under the assumption that AEs within a SOC are more similar to each other than across a SOC. This approach directly models the biological relationship among different AEs. The Berry and Berry method is expanded to the Poisson model to account for different follow-up times in clinical trials.[17] The performances of different methods via simulation are compared, with the conclusion that the Bayesian hierarchical modeling approach outperforms the other methods, including the unadjusted Fisher exact test, the Benjamini-Hochberg approach,[18] and the double false discovery rate method.[15] Some practical considerations in implementing these Bayesian hierarchical modeling methods for safety signal detection in clinical trials are provided.[5] In addition to Bayesian hierarchical modeling of AEs' structure, a Bayesian screening approach for identifying AEs that may reflect real toxicities has been proposed.[19] The approach is based on the posterior probability that the same process generated the event rates in both control and treatment groups in determining whether the relationship between a testing drug and an AE warrants further investigation. The false-positive rate of this process is not inflated by multiplicity. Furthermore, the diagnostic properties of the method (ie, the usual diagnostic probabilities including false-positive and -negative rates, positive and negative predictive values) can be obtained explicitly. In addition, an analysis method for safety data from a pool of clinical studies called multivariate Bayesian logistic regression (MBLR) has been introduced.[20] Essentially, MBLR allows information from the different issues to "borrow strength" from each other so that the method is especially suited for sparse event data. The method enables a search for vulnerable subgroups based on the covariates in the regression model. The method requires the selection of a set of medically related issues, potentially exchangeable with respect to their dependence on the treatment and covariates.

In analyzing postmarketing spontaneous reports, 2 standard disproportionality (DP) analysis methods have been developed. A Bayesian confidence propagation neural network (BCPNN) approach was utilized to analyze a World Health Organization (WHO) database of adverse drug reactions.[21] A gamma-Poisson shrinkage algorithm was proposed to analyze the FDA's AERS database.[22] The Bayesian shrinkage concept is intrinsic to both methods in attempting to deal with small observed counts and associated instability. The impact of stratification in standard DP analyses of spontaneous reporting databases is evaluated by introducing the notion of "overstratification" and showing that selective stratification can modestly improve the performance of signal detection.[23] A white paper on behalf of the PhRMA-FDA Collaborative Working Group was published to provide an overview of currently used safety data-mining methods for spontaneous reporting databases, their strengths and limitations, as well as analytical considerations for using these methods and interpreting the results.[24]

The development of statistical methods for signal detection in electronic medical records (EMRs) and administrative claims databases remains an important challenge. The Observational Medical Outcomes Partnership (OMOP) Cup was a recent public competition for signal detection methods, aiming to accurately classify the drug-event pairs in longitudinal EMR and claims databases. The top-performing method for OMOP Cup 2009-2010[25] is an adaptation from methods used for spontaneous signal detection of AE data to longitudinal observational databases and has 2 steps: (1) longitudinal gamma-Poisson shrinker (LGPS) is used to identify potential signals, and (2) longitudinal evaluation of observational profiles of AEs related to drugs (LEOPARD) is used to remove protopathic bias. The paper showed that the performance of LGPS and LEOPARD was better than all other methods entered in the competition, including Bayesian logistic regression and BCPNN. Another method, called the temporal pattern discovery approach, uses Bayesian shrinkage to again protect spurious associations, contrasts event rates in different periods to filter out indications for treatment, and proposes a graphic approach to characterize temporal patterns and facilitate clinical interpretation.[26,27]

Meta-analysis for Analyzing Safety Data

Meta-analysis techniques have been increasingly utilized to identify and evaluate safety issues in drug development. Very often, a meta-analysis of randomized controlled trials (RCTs)

Table 1. Advantages and disadvantages in choosing between an absolute measure versus a relative measure.

	Advantages	Disadvantages
Absolute measure	• Easy to interpret • Always well defined so it allows the inclusion of studies with zero events • Knowledge of absolute risks is important in clinical decision making	• Clinical importance may depend on the underlying baseline event rate, but it is less of an issue for rare events
Relative measure	• Typically analyzed on a logarithm scale; more stable on average than absolute measures • Good statistical properties	• Undefined when the control rate is zero, so it does not allow the inclusion of studies with zero events

is used as a way to improve power in assessing rare events. In recent years, we have seen a few highly visible cases in which a meta-analysis of RCTs was used to evaluate the safety profile of certain drugs, ultimately leading to key regulatory decisions.[28,29]

However, the conduct of a meta-analysis in this context poses some challenges, both in general and on statistical grounds. In general, a meta-analysis is both a type of medical research and a statistical approach. A poorly conceived meta-analysis design cannot be rescued by even the highest quality statistical meta-analysis. The ICH Statistical Principles for Clinical Trials E9 (ICH-E9)[30] and the SPERT[3] stress that a meta-analysis should be prospectively planned with the clinical trials program in the development of a new treatment rather than post hoc. It is very important to define a precise question to address with the meta-analysis, and a well-planned meta-analysis protocol and statistical analysis plan should be created up front.

As the most important analytical principle for analyzing data from multiple studies, the meta-analytic technique (ie, stratification by study) should be used. Simply pooling data across studies without stratification should be avoided, as it may suffer from the issues related to "Simpson's paradox," especially when studies included have different randomization ratios.[31,32]

A meta-analysis based on RCTs is a powerful tool but poses a series of methodological challenges that require due attention and action. Although there are many other related issues in this context such as outcome ascertainment, data quality, sensitivity analyses, and clear and transparent reporting, for the purpose of this paper, we focus on a few important statistical considerations.

Scale of Measures

When choosing a scale for an overall treatment effect in a meta-analysis, it is recommended that one consider the consistency of the effect across studies, the mathematical properties, and ease of interpretation.[33,34] There are advantages and disadvantages to choosing between absolute measures (eg, risk difference) and relative measures (eg, odds ratio or relative risk) (Table 1) for a dichotomous outcome. A meta-analysis is most often conducted on a relative scale. However, the absolute difference may be appealing for rare events because it is readily interpretable and facilitates the inclusion of studies with no events. When considering effect measures across studies within a meta-analysis, risk differences tend to show more heterogeneous results than relative measures do.[35] One recommendation is to use an odds ratio to produce a meaningful overall estimate and then convert it to a risk difference to help with the clinical or public health interpretation.[36,37] When the studies in a meta-analysis involve time-to-event data, the most appropriate statistics are the logarithm of HR and its variance. When such statistics are missing in a publication or study report, a number of methods exist for estimating these statistics for a variety of situations.[38,39]

Fixed-Effect Versus Random-Effects Models

There are 2 commonly used statistical models for a meta-analysis: the fixed-effect and random-effects models. The fixed-effect model assumes a common treatment effect across studies. By contrast, the random-effects model allows that the underlying true treatment effects may vary from study to study.[40] The CIs for the summary effect are wider under the random-effects model than under the fixed-effect model. When the studies are homogeneous, the 2 models yield similar results. Random-effects models gained wider acceptance recently because they take into account the between-study heterogeneity. However, in the meta-analysis setting of rare events, the use of random-effects models is controversial because estimates for between-study heterogeneity are often extremely unstable and therefore may be misleading. Some authors[41] advocate that the fixed-effect model is preferred in the meta-analysis of rare events, while others recommend it might be generally useful to conduct a meta-analysis with both models for comparison. Results that vary substantially between these

2 approaches should be examined carefully to understand the clinical reasons behind the observed differences.[32] Some analytical approaches such as stratification by patient characteristics, models of individual-level data using treatment by covariate interactions, and meta-regression methods may be helpful to understand the heterogeneity.

Heterogeneity Assessment

Assessing heterogeneity, both qualitatively and quantitatively, is an essential part of a meta-analysis. Heterogeneity can be classified into 3 categories: clinical, methodological, and statistical heterogeneity.[36,42] Statistical heterogeneity is evaluated graphically with forest plots[43] or through a heterogeneity test that examines the null hypothesis that all studies are evaluating the same effect. A common approach to assess heterogeneity is to use the Cochran Q statistic with the χ^2 test. If the test result is significant, it indicates that the results are heterogeneous, and an overall effect estimate may not be a good representation of the results across studies. However, this test is known to have low power when the number of studies is small. The strategy of starting with a fixed-effect model and then moving to a random-effects model if the test results of heterogeneity are significant is flawed and should be strongly discouraged.[40] Furthermore, one does not need a significant global test or estimate of heterogeneity to justify a prespecified exploration of treatment effect modification. Absence of a nonsignificant Q statistic does not necessarily mean absence of a treatment effect modifier. Higgins and Thompson[44] advocated the quantification of heterogeneity using the I^2 statistic, which describes the percentage of total variation across studies that is due to heterogeneity rather than chance. The advantage of I^2 is that it can be directly compared between meta-analyses with different numbers of studies, types of outcome data, and choice of effect measure. I^2 has been commonly used to help readers assess the consistency of meta-analysis results across studies.[45] However, even I^2 has its challenges. If the component studies are very large, then one can obtain a large value for I^2 because the within-study variability is small, even if the among-study variability is small from a clinical perspective.

Statistical Methods for Meta-analyses of Rare Events

Rare events pose some unique analytic challenges for meta-analysts. Standard inferences for a meta-analysis rely on large sample approximations. They may not be accurate and reliable when sample sizes from individual studies are small, when the total number of studies is small, and when the total number of AEs is small. Some serious AEs are often sparse, leading to zero events being observed in one arm or even both arms for some studies. The problem with lack of power in evaluating heterogeneity is amplified when the number of studies is only modest and an event of interest is rare.

Preferred methods for rare events have been addressed in a few recent papers.[33,41,46] Bradburn et al[46] evaluated the performance of different methods for binary outcomes. At event rates below 1%, the Peto method provides the least biased, most powerful estimate and best CI coverage for balanced groups, but bias increases with greater group imbalance and larger treatment effects. The Mantel-Haenszel method performs well under many circumstances. Sweeting and colleagues[41] recommended an alternative continuity correction when there are zero events: namely, continuity correction of the ''treatment arm'' (which is based on the reciprocal of the opposite group size) rather than the usual method of adding 0.5 to all cells of a 2×2 table when 1 cell contains a zero. Both inverse variance–weighted averages and the DerSimonian and Laird methods should be avoided in the setting of rare events given their known poor performance. Logistic regression and the Bayesian fixed-effect model perform consistently well, irrespective of group imbalance.

A recent advancement in the meta-analysis of rare events is the exact inference procedure, which includes zero event studies.[47] An unconditional approach (by including zero event studies) based on the Poisson random-effects model in the rare event setting also has been proposed.[48] In addition, Bayesian methods can be appropriately applied to a meta-analysis of rare events in which the use of hierarchical models can modulate the extremes in the zero event setting, borrowing information from studies with events to derive posterior inferences for the treatment effect estimates. Furthermore, Bayesian methods can deal with complex modeling. Askling et al[49] used the Bayesian hierarchical piecewise exponential survival model to investigate the cancer risk for the drug class of tumor necrosis factor inhibitors. The Bayesian model is able to analyze the individual patient-level meta-data, taking into account patient-level and possibly time-dependent covariates and models between study heterogeneity. Kaizar et al[50] used the Bayesian hierarchical model to quantify the risk of suicidality in children who use antidepressants.

Meta-analysis of Individual Patient Data (IPD)

Although many published meta-analyses are based on aggregate trial-level summary data, often retrieved from the published literature, there is growing attention on meta-analyses based on IPD. Access to patient-level data provides greater flexibility and is generally superior to summary-level information, although IPD are not always available. A meta-analysis of IPD is of importance for the pharmaceutical industry as companies often have all the patient-level data in their clinical trial database through the product's life cycle. It is especially true before a product receives marketing authorization. A meta-

analysis of IPD offers many advantages: (1) it enables analysts to use common definitions, coding, and cut points and produces consistent analyses across multiple studies; (2) it facilitates the exploration of heterogeneity at the patient level and subgroup analyses of patient-level data; (3) it permits the investigation of additional hypotheses (particularly those related to individual patient characteristics) in which the data would be unavailable in published results; (4) it allows adjustments for the same covariates across studies; and (5) it permits analyses of time to events and allows analysts to address long-term outcomes when analyzing events with long latency. This is vital when dealing with nonconstant or nonproportional hazards. For these reasons, a meta-analysis of IPD is regarded as the gold standard and, when feasible, should be considered.[37,51]

Multiplicity

Multiplicity is a challenging and controversial issue in meta-analyses for drug safety evaluation, where adjustment for multiplicity is not commonly performed, and there is no consensus in the scientific community. Berlin et al[37] gave an overview of the issues related to multiple looks and/or multiple endpoints in this context. Whether to adjust for different types of multiplicity should tie with the analytical goals. As described in section 4, for tier 1 events, since we know the hypothesis in advance and the analytical goal is to quantify the risk, the adjustment of multiplicity in the setting of multiple looks or cumulative meta-analyses warrants considerations. Some caution may be advised in interpreting cumulative meta-analyses for the increased probability of a spurious positive finding introduced by the use of repeated statistical tests.[52] In light of the FDA guidance on evaluating the CV risk for type 2 diabetes, Ibrahim et al[53] developed a Bayesian meta-analytic sample size determination method for planning a phase 2/3 antidiabetic drug development program to ensure good operating characteristics of the program's design in meeting specific criteria for type I error and power. Chen et al[54] extended the previous work by developing a novel Bayesian sequential meta-experimental design approach to address the multiplicity issue in the context of sequential meta-analyses. For non–tier 1 events, since the analytical goal is signal detection, the multiplicity consideration should focus on adjustment for multiple endpoints. The Bayesian signal detection techniques aforementioned in section 4 could be considered by further extending to the meta-analysis setting.

Safety Graphics

Effective and clear presentation of safety information is crucial for safety evaluation and communication. Numerous big tables and long listings could obscure the true safety signal detection. On the other hand, visualization of data and results can provide an effective presentation of complex data and facilitate the signal detection and decision-making process. Consistent with the recommendations of the SPERT, Chuang-Stein and Xia[5] presented a comprehensive summary of graphic displays in the safety evaluation of clinical trials.

To facilitate a palette of graphics for safety data visualization, and to share best practices for statistical graphics, the FDA/Industry/Academia Safety Graphics Working Group was formed in 2009. The group has approximately 20 statisticians from the FDA, industry, and academia. Not only the graphics but also brief descriptions of the graphics, datasets, and sample codes used to create the graphs are provided in a publicly available repository. It has recommendations on the effective use of graphics for 3 key safety areas: AEs, ECGs, and laboratory analytes, along with good principles (see the FDA/Industry/Academia Safety Graphics wiki site: http://www.ctspedia.org/do/view/CTSpedia/WorkingGroupInformation). The working group focused on static graphs; interactive capabilities could be beneficial but are considered out of the scope of the working group's efforts.

The FDA/Industry/Academia Safety Graphics Working Group believes that it is important to come up with clinical questions first and then construct the appropriate analyses to address those clinical questions. For example, for AEs, there are 6 questions that are critical for safety evaluation:

- Which AEs are elevated in treatment versus control?
- Which AE could be a safety signal?
- Is there a difference in time to the first event across treatment groups?
- What are the trends of time to the first event among different AEs?
- Which AEs are elevated in patient subgroups?
- What are the risk factors for an AE?

To address those questions, the wiki site presented suggested plots such as a volcano plot and a time-to-AE occurrence plot, along with datasets and codes. To illustrate the working group's efforts, Anziano and Gordon[55] described the utilization of statistical graphics in areas of cardiac safety and hepatotoxicity, which are critical in safety evaluation. In addition, Duke et al[56] looked specifically at elements of good graphics; common safety questions and recommended graphics in the areas of AEs, ECGs and vital signs, and laboratory parameters (especially liver toxicity); guidance from regulatory agencies; and the future of graphics in benefits and risks and interactive graphics.

Visualization is an effective tool not only to present safety data but also the B-R profile. The Innovative Medicine Initiative Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium (IMI PROTECT) performed a comprehensive review of graphics and made a set of recommendations.[57] Other organizations are trying to

recommend graphics for B-R assessments as well.[58] Considering the easy implementation and communication of these B-R graphics in industry, graphics such as a forest plot, bar graph, line graph, tree diagram, stacked bar chart, and difference display could provide an effective presentation of complex data to a variety of stakeholders, enhance transparency, and improve the ability to make decisions.[59] For a forest plot, if the measurement metrics (eg, relative risk, risk difference, HR) are different for different endpoints, the data and results can be presented in different graphics panels as well. In addition, if needed, such forest plot graphics can be generated for multiple endpoints in an individual study, subgroups in an individual study, and meta-analyses to facilitate the presentation and interpretation of benefits and risks.

B-R Assessment

Evaluations of both safety and efficacy are critical during drug development.[60] While the evaluation of new treatments has always involved a B-R assessment, these assessments have tended to be informal, somewhat subjective, and lacking in transparency. Increasingly, companies, regulatory agencies, and other governing bodies are taking a more systematic approach and are beginning to use structured B-R assessment approaches.

In fact, B-R assessments are very important. They can account for relative benefits and risks, allow stakeholders to continuously determine whether products should be approved or reimbursed for proposed or approved indications, and support the clear communication of a product's attributes between sponsor and stakeholder groups. The B-R assessment requires an evaluation of efficacy, safety, and quality and considers the disease condition, unmet medical needs, benefits, risks, and risk management. It is important to take into account the patient perspectives as well. The drug benefits need to outweigh the risks through the drug's life cycle.

A number of organizations and initiatives have been actively investigating B-R assessments[61-64] including the following:

- In the US, the FDA, as part of the Prescription Drug User Fee Act (PDUFA) V negotiations completed in 2012, has agreed to hold a series of disease-specific public meetings to obtain broad input on patient outcomes and strategies of importance. In addition, the FDA is committed to a series of meetings and workshops during 2013-2018 to develop a B-R framework. The FDA also drafted the PDUFA V Implementation Plan on Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making. A framework by the PhRMA Benefit-Risk Action Team (BRAT) was introduced as well.[65] This initiative was transitioned to the Centre for Innovative Regulatory Sciences (CIRS) in 2012 for broadened input and further development.
- In Europe, the EMA's reflection paper on benefits and risks stated that expert judgment is expected to remain the cornerstone of B-R evaluations and that quantitative B-R assessments are not expected to replace qualitative evaluations.[66] The EU's pharmacovigilance legislation placed increased emphasis on ongoing B-R assessments with the Periodic Benefit-Risk Evaluation Report.[67] Now, the Periodic Safety Update Report includes subsections for risk evaluation (already in the template), benefit evaluation, and B-R assessment (which are both new to the template).
- In other regions such as Canada, there is an increasing emphasis to incorporate B-R assessments into approval and marketing decisions as well.

There is a large body of literature on B-R methods; many methods have been proposed, and several recommendations have been made,[68,69] both descriptive and quantitative, on how to assess and weigh benefits and risks, including recommendations from the IMI PROTECT, International Society for Pharmacoeconomics & Outcomes (ISPOR), PhRMA BRAT, and CIRS–Unified Methodologies in Benefit-Risk Assessment (UMBRA) Initiative. In the US, the Quantitative Sciences in the Pharmaceutical Industry Benefit-Risk Working Group (QSPI BRWG) has been formed among statisticians to work on critical issues on benefits and risks.

While formal approaches for B-R assessments are still evolving, there are quite a few proposed frameworks, including the framework from the BRAT; Problems, Objectives, Alternatives, Consequences, Trade-offs, Uncertainty, Risk attitudes, and Linked decisions (PrOACT-URL); FDA; or CIRS. Frameworks for B-R assessments provide a structured approach in assessing the product's B-R profile and are the first step prior to the application of quantitative statistical methodologies. The content that populates the frameworks may be either descriptive, quantitative, or both. To this point, B-R assessments are largely based on a qualitative approach. Quantitative approaches such as the multicriteria decision analysis can be used in more complex situations.

The number of patients needed to be exposed to a treatment to cause harm in a patient (NNH) and the number of patients needed to be treated for one to benefit (NNT) have been used quite often despite their limitations. One limitation of this approach is that it is designed for the case of a single benefit and a single harm, which rarely occurs. Another limitation of using NNT and NNH is generating CIs when the CI for the risk difference includes zero. A further problem arises in attempting to apply these statistics to outcomes over time; different values of the statistics would be obtained at different time periods. To address this

concern, one could use an exposure-adjusted approach that assesses the benefit and harm as a rate per unit time.

Thus, a B-R assessment remains challenging for a number of reasons, including but not limited to the following[68]:

- First, the lack of clarity as to how regulators weigh B-R information in approval decisions and the lack of standardized and validated methods to weigh and quantify benefits and risks pose challenges.
- Second, a B-R assessment needs to take into account the underlying disease setting and the unmet medical needs. However, perspectives can differ regarding the relative importance of specific benefits and harms.
- Third, it may not be clear which endpoints should be included, and there is often the potential for an unbalanced B-R evaluation. Correlations may exist among endpoints (eg, between progression-free survival and overall survival). Therefore, B-R assessments need to properly account for the correlation to avoid inflating benefits or risks. In addition, a B-R evaluation could be unbalanced. Pivotal trials tend to focus on the evaluation of benefits and collect detailed information on efficacy variables. This may include factors such as training investigators on the assessment of these endpoints and endpoint adjudication committees. There is typically less emphasis on safety variables. In long-term follow-up and postmarketing studies, on the other hand, the focus is typically on safety, and there may not be sufficient information on benefits. Excluding a benefit (or risk) from an evaluation is the same as giving it a weight of zero.
- Furthermore, a challenge is how to select the weights while accounting for patient preference. Weighting is key for B-R approaches in combining benefits and risks.[70] A few B-R approaches assume equal weighting of benefits and risks, which often is not appropriate. Weight selection is somewhat subjective and is a matter of clinical judgment. Weight selection could be different for different disease areas and could depend on input from patients, regulatory agencies, sponsors, and payers. However, making selected weights clear increases transparency. It is helpful to standardize weights or methods to solicit weights for specific indications, if feasible. Of course, it is helpful to conduct sensitivity analyses to get a sense for the robustness of the conclusion to different weights. Note that patient perspectives are important to consider when selecting weights.
- Finally, global harmonization is critical and needs to be coordinated as well.

In summary, the B-R landscape is very much evolving. There are increased interests and efforts in further enhancing structured B-R assessments. There are currently no commonly accepted B-R methodologies. Several methodologies are being tested as part of ongoing initiatives. Choosing one approach for every decision problem may not be realistic because each has its own strengths and weaknesses, and sometimes its pragmatic applications are limited by available evidence and underlying assumptions and most often are limited by the resources and the ability to effectively communicate the results from a B-R analysis. A B-R assessment is very important to improve transparency and communication but is complicated as well. Cross-functional efforts and close collaboration are critical.

Concluding Remarks

In this paper, we have highlighted many of the challenges associated with safety evaluations and provided a review for key areas in which statistical methods can be leveraged to enhance the analysis, reporting, and interpretation of safety assessment. Analytical approaches in the safety area are generally not sufficiently sophisticated, but the research is beginning to take off. A PSAP can greatly facilitate safety evaluation and signal detection. A B-R assessment needs to be conducted systematically through management of the drug's life cycle. We encourage statisticians to become more involved early in the drug development process in safety evaluations and B-R assessments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

References

1. ICH. ICH safety guidelines. Available at: http://www.ich.org/products/guidelines/safety/article/safetyguidelines.html. Accessed October 29, 2013.
2. Council for International Organizations of Medical Sciences (CIOMS) Working Group VI. *Management of Safety Information From Clinical Trials*. Geneva: CIOMS; 2005.
3. European Commission. Communication from the commission: detailed guidance on the collection, verification and presentation of adverse event/reaction reports arising from clinical trials on medicinal products for human use ('CT-3'). 2011. Available at: http://www.kme-nmec.si/Docu/ct-3.pdf. Accessed October 29, 2013.
4. 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 Trials*. 2010;6(5):430-440.
5. Chuang-Stein C, Xia HA. The practice of pre-marketing safety assessment in drug development. *J Biopharm Stat*. 2013;23(1):3-25.

6. Xia HA, Crowe BJ, Schriver RC, Oster M, Hall DB. Planning and core analyses for periodic aggregate safety data reviews. *Clin Trials*. 2011;8(2):175-182.
7. Downs M, Wittes J. Data monitoring in practice: making your data monitoring committee effective. ASA Biopharmaceutical Section Webinar. March 10, 2010. Available at: http://www.biopharmnet.com/doc/2010_03_10_webinar.pdf. Accessed October 29, 2013.
8. O'Neill RT. Assessment of safety. In: Peace KE, ed. *Biopharmaceutical Statistics for Drug Development*. New York: Marcel Dekker; 1988:543-604.
9. Solomon SD, McMurray JJ, Pfeffer MA, et al. Cardiovascular risk associated with celecoxib in a clinical trial for colorectal adenoma prevention. *N Engl J Med*. 2005;352:1071-1080.
10. Proschan MA, Lan KKG, Wittes JT. *Statistical Methods for Monitoring Clinical Trials*. New York: Springer; 2006.
11. MedDRA. MSSO introductory guide: version 16.0. 2013. Available at: http://www.meddra.org/sites/default/files/guidance/file/intguide_16_0_english.pdf. Accessed October 29, 2013.
12. Gould AL.Drug safety evaluation in and after clinical trials. Presented at: the Deming Conference; December 3, 2002; Atlanta City, New Jersey.
13. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, Florida: Chapman and Hall/CRC Press; 2010.
14. Prieto-Merino D, Quartey G, Wang J, Kim J. Why a Bayesian approach to safety analysis in pharmacovigilance is important. *Pharm Stat*. 2011;10(6):554-559.
15. Mehrotra DV, Heyse JF. Use of the false discovery rate for evaluating clinical safety data. *Stat Methods Med Res*. 2004;13:227-238.
16. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*. 2004;60:418-426.
17. Xia HA, Ma H, Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *J Biopharm Stat*. 2011;21(5):1006-1029.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.
19. Gould AL. Detecting potential safety issues in clinical trials by Bayesian screening. *Biom J*. 2008;50:837-851.
20. DuMouchel W. Multivariate Bayesian logistic regression for analysis of clinical study safety issues. *Stat Sci*. 2011;27(3):319-339.
21. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*. 1998;54:315-321.
22. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat*. 1999;53:177-190.
23. Hopstadius J, Norén GK, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. *Drug Saf*. 2008;31(11):1035-1048.
24. Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Saf*. 2005;28:981-1007.
25. Schuemie MJ.Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Drug Saf*. 2011;20:292-299.
26. Norén G, Hopstadius J, Bate A, Star K, Edwards I. Temporal pattern discovery in longitudinal electronic patient records. *Data Min Knowl Discov*. 2010;20(3):361-387.
27. Norén GN, Bate A, Hopstadius J, Edwards IR. Safety surveillance of longitudinal databases: methodological considerations. *Pharmacoepidemiol Drug Saf*. 2011;20:714-717.
28. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*. 2007;356(24):2457-2471.
29. Bridge JA, Iyengar S, Salary CB, et al. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: a metaanalysis of randomized controlled trials. *JAMA*. 2007;297(15):1683-1696.
30. ICH. ICH Harmonised Tripartite Guideline: statistical principles for clinical trials E9. 1998. Available at: http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html. Accessed May 23, 2013.
31. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ*. 1997;315:1533-1537.
32. Hammad TA, Pinheiro SP, Neyarapally GA. Secondary use of randomized controlled trials to evaluate drug safety: a review of methodological considerations. *Clin Trials*. 2011;8(5):559-570.
33. Sutton AJ, Cooper NJ, Lambert PC, Jones DR, Abrams KR, Sweeting MJ. Meta-analysis of rare and adverse event data. *Expert Rev Pharmacoecon Outcomes Res*. 2002;2:367-379.
34. Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Smith GD, Altman DG, eds. *Systematic Reviews in Health Care: Meta-analysis in Context*. 2nd ed. London: BMJ Publishing Group; 2008:313-335.
35. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med*. 2000;19(13):1707-1728.
36. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *J Clin Epidemiol*. 2007;60(9):874-882.
37. 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 Trials*. 2013;10(1):20-31.
38. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med*. 1998;17:2815-2834.
39. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials*. 2007;8:16.
40. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random effects models for meta-analysis. *Res Synth Methods*. 2010;1:97-111.
41. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004;23:1351-1375.

42. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*. 1994;309:1351-1355.
43. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ*. 2001;322:1479-1480.
44. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539-1558.
45. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions: Version 5.1.0. Cochrane Collaborations. March 2011. Available at: http://handbook.cochrane.org/. Accessed May 28, 2013.
46. Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26:53-77.
47. Tian L, Cai T, Pfeffer M, Piankov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta analysis and its application to the analysis of independent 2 by 2 tables with all available data but without artificial continuity correction. *Biostatistics*. 2009;10(2):275-281.
48. Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Stat Med*. 2010;29(20):2078-2089.
49. Askling J, Fahrbach K, Nordstrom B, Ross S, Schmid C, Symmons D. Cancer risk with tumor necrosis factor alpha (TNF) inhibitors: meta-analysis of randomized controlled trials of adalimumab, etanercept, and infliximab using patient level data. *Pharmacoepidemiol Drug Saf*. 2011;20(2):119-130.
50. Kaizar EE, Greenhouse JB, Seltman H, Kelleher K. Do antidepressants cause suicidality in children? A Bayesian meta-analysis. *Clin Trials*. 2006;3(2):73-90, discussion 91-98.
51. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Med Res Methodol*. 2005;5:14.
52. Hu M, Cappelleri JC, Lan KK. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clin Trials*. 2007;4(4):329-340.
53. Ibrahim J, Chen M, Xia HA, Liu T. Bayesian meta-experimental design: evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics*. 2012;68(2):578-586.
54. Chen M, Ibrahim J, Xia HA, Liu T, Hennessey V. Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program. *Stat Med*. Provisionally accepted.
55. Anziano R, Gordon R. Examples of understanding safety data using graphical methods. *Wiley Interdiscip Rev Comput Stat*. 2013;5:78-95.
56. Duke S, Jiang Q, Huang L, Banach M. Safety graphics. In: Jiang Q, Xia HA, eds. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. New York: Taylor & Francis. In preparation.
57. Mt-Isa S, Peters R, Phillips LD, et al. On behalf of PROTECT Work Package 5 participants: review of visualisation methods for the representation of benefit risk assessment of medication. February 2013. Available at: http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBRassessmentofmedicationStage1F.pdf. Accessed October 27, 2013.
58. Zeng D, Chen M, Ibrahim JG, et al. A counterfactual p-value approach for benefit-risk assessment in clinical trials. *J Biopharm Stat*. In press.
59. Jiang Q, Shepherd S, Ke C, Ma H, Snapinn S. Considerations for benefit: risk assessment in pharmaceutical drug development. Presented at: ICSA/ISBS; June 9-12, 2013; Bethesda, Maryland.
60. Chuang-Stein C. A new proposal for benefit-less-risk analysis in clinical trials. *Control Clin Trials*. 1994;15:30-43.
61. EMA. Human Medicines Development and Evaluation. Benefit-Risk Methodology Project. Work package 2 report: applicability of current tools and processes for regulatory benefit-risk assessment. August 31, 2010 (EMA/549682/2010). Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/10/WC500097750.pdf. Accessed October 27, 2013.
62. EMA. Human Medicines Development and Evaluation. Benefit-Risk Methodology Project. Work package 4 report: benefit-risk tools and processes. May 9, 2012 (EMA/297405/2012). Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2012/03/WC500123819.pdf. Accessed October 27, 2013.
63. FDA. Structured approach to benefit-risk assessment in drug regulatory decision-making. Draft PDUFA V Implementation Plan: February 2013, Fiscal Years 2013-2017. Available at: http://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm329758.pdf. Accessed October 27, 2013.
64. Walker S.Chairman's introduction. Presented at: CIRS Technical Workshop on Benefit-Risk Framework for the Assessment of Medicines: Valuing the Options and Determining the Relative Importance (Weighting) of Benefit and Risk Parameters; December 13, 2012; Philadelphia, Pennsylvania.
65. Levitan BS, Andrews EB, Gilsenan A, et al. Application of the BRAT framework to case studies: observations and insights. *Clin Pharmacol Ther*. 2011;89(2):217-224.
66. EMA. Report of the CHMP Working Group on benefit-risk assessment models and methods. January 19, 2007 (EMEA/CHMP/15404/2007). Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedura l_guideline/2010/01/WC500069668.pdf. Accessed October 27, 2013.
67. EMA. ICH guideline E2C (R2) on periodic benefit-risk evaluation report (PBRER). January 2013 (EMA/CHMP/ICH/544553/1998). Available at: http://www.rsihata.com/updateguidance/2013/WC500136402.pdf. Accessed October 27, 2013.
68. Jiang Q, Ke C, Bloss L, Kouchakji E, Snapinn S, Vega J. Some thoughts on benefit-risk assessment. Presented at: FDA/Industry Statistical Workshop; September 16-18, 2013; Washington, DC.
69. Ke C, Jiang Q, Snapinn S. Methods for benefit-risk assessments in drug development. In: Jiang Q, Xia HA, eds. *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. New York: Taylor & Francis. In preparation.
70. Kouchakji E, Jiang Q, Consuelo B, Bloss L, Ke C. Benefit:risk assessments: a ranking/qualitative approach. Presented at: CIRS Technical Workshop on Benefit-Risk Framework for the Assessment of Medicines: Valuing the Options and Determining the Relative Importance (Weighting) of Benefit and Risk Parameters; December 13, 2012; Philadelphia, Pennsylvania.