Medical Safety & Global Health
ISSN: 2574-0407

Miao et al., Med Saf Glob Health 2017, 6:1
DOI: 10.4172/2574-0407.1000133

Commentary Open Access

Sensitivity *versus* Specificity in the Evaluation of Adverse Event Data from Clinical Trial

Miao J*, Lai TL, Chen J and Heyse JF

Department of Statistics, Stanford University, CA, USA

Abstract

The evaluation of safety is an important part of clinical trials of pharmaceutical, biological, and vaccine products. In early phase trials, the evaluation is mostly exploratory with a focus primarily on serious adverse reactions to the candidate product. In later phases of clinical development programs the safety profile is characterized more fully using larger numbers of patients. Unlike the evaluation of drug efficacy, the outcome of which is based on a single or a collection of prespecific hypotheses, the hypotheses to test to conclude a drug has potential safety burden is generally not prespecified. The test and conclusion of potential safety issue of a drug are usually based on an arbitrary number of reports of adverse events that have not been identified at the outset, which amounts to using observed data to test hypotheses that are generated by the same data.

Keywords: Drug safety; Clinical trial; Observational study; Double False Discovery Rate (DFDR); Multiple hypotheses testing

Introduction

The collection of safety and tolerability data in clinical trials goes well beyond the data collected to address specific safety hypotheses, which may be developed from the chemical or biological properties of the product, or possibly from observations from early-phase non-clinical and clinical trials. Adding to the complexity, the set of possible adverse effects is very large and new unanticipated effects are always possible. Moreover, confirmatory clinical trials to test the efficacy hypotheses usually have large sample sizes, and this may result in many more adverse event types, some of which were not expected based on the pharmacological profile of the product, preclinical experiments in animals, or *in vitro* studies. Hence there is potential for drawing false positive conclusions and the need for understanding the multiplicity aspects in safety signal detection. Safety assessment continues into the post-marketing phase with clinical trials in which specific safety issues may be addressed, and with post-marketing surveillance and pharmacovigilance plans that are usually based on large databases of patient electronic medical records and spontaneous reports of adverse events. While the multiplicity considerations differ during different phases of drug development, they are always an important component in the analysis and interpretation of clinical safety data. In their discussion of safety analysis in the pre-licensure phases, Xia et al. [1] and Chuang-Stein and Xia [2] identify multiplicity as a key issue that needs to be included in the clinical development plan for a new medical product. Since almost all clinical trials are designed with the objective of evaluating a product's efficacy for its regulatory approval, the study design, endpoint selection, and sample size determination are usually based on the efficacy hypothesis. For safety, there is often no specific hypothesis to test in the clinical trial design, but the study plan still collects and analyses adverse experiences reported by the study participants. Adverse event data should be carefully catalogued and summarized using standard coding dictionaries such as MedDRA (Medical Dictionary for Regulatory Activity). Crowe et al. [3] have pointed out the potential for too many false positive safety signals if the multiplicity problem is ignored. Kaplan et al. [4] give an example of how false positive signals can impact the interpretation of the safety profile of the drug or vaccine. This example is about a safety and immunogenicity trial to compare a combination vaccine, labelled A, to one of its individual component vaccines, labelled B, in an infant population. The analysis of the adverse event data identifies UHPC (Unusual High Pitched Crying) as the single event with an individual P-value<0.05; the incidence of UHPC for group A was 6.7% compared to 2.3% for group B, yielding a two-sided P-value of 0.016. However, UHPC was just one of 92 adverse experience types in the study, and there was no medical rationale for this finding, nor were there additional data suggesting such a relationship from the already approved and marketed components of the combination vaccine. To address the multiplicity issue, the study team undertook a confirmatory study requested by regulators. The large follow-up trial concluded that the original P-value, unadjusted for multiplicity, was a false positive signal. Hence a significant amount of time and money was expended on chasing down what could easily have been determined to be not statistically significant by using appropriate multiplicity adjustments in the original analysis.

There is an implicit trade-off between sensitivity and specificity in the evaluation of clinical safety data. The preceding paragraph and the references cited therein are related to specificity, which is the proportion of true negative effects correctly identified as such by the safety evaluation. Thus, 1-specificity is the aforementioned false positive rate, which corresponds to the type I error in hypothesis testing. Sensitivity is the proportion of true positive effects correctly identified as such by the safety evaluation and corresponds to "power", or 1-type II error, in hypothesis testing. The issue here arises from a very large number of hypotheses, many of which may not be specified in advance. This commentary is on some approaches to the treatment of this issue and the extent to which they address the trade-off between sensitivity and specificity.

***Corresponding author:** Jing Miao, Department of Statistics, Stanford University, CA 94305, USA, Tel: +1650 7232300; E-mail: miaojing1993@gmail.com

Received May 11, 2017; **Accepted** May 22, 2017; **Published** May 30, 2017

Citation: Miao J, Lai TL, Chen J, Heyse JF (2017) Sensitivity *versus* Specificity in the Evaluation of Adverse Event Data from Clinical Trial. Med Sif Glob Health 6: 133. doi: 10.4172/2574-0407/1000133

MedDRA Categorization of Adverse Events and Data Tabulation

Mehrotra and Heyse [5] were the first to (a) draw attention to the multiplicity issue in safety evaluation of clinical trials data and (b) propose a method, called Double False Discovery Rate (DFDR) control, to address this issue. They consider the adverse event data from a safety and immunogenicity trial of a measles, mumps, rubella, varicella (MMRV) combination vaccine trial. The study population included healthy toddlers, 12-18 months of age. The comparison of interest was between Group 1: MMRV+PedvaxHIB on Day 0, and Group 2: MMR+PedvaxHIB on Day 0, followed by an optional varicella vaccination of Day 42. The safety follow-up included local and systemic reactions over Days 0-42 for N=148 in Group 1, and N=132 in Group 2 over Days 42-84. The follow-up duration of 42 days is standard for live virus vaccines such as varicella. The question, which involves the varicella component of MMRV, is whether the safety profile differs between its administration in a combination and giving it 6 weeks later as a monovalent vaccine. The adverse events are coded using a standard dictionary (e.g., MedDRA) and classified into groupings by body systems. The MMRV dataset consists of 40 adverse event types which are categorized into 8 body systems, as shown in the first three columns of Table 1, in which b represents the body system index, and i the index of adverse event types within a certain body system.

We next give some background about these body system groupings in adverse event dictionaries such as MedDRA, which is a hierarchically structured vocabulary (http://www.meddra.org/). MedDRA's five-level hierarchy of terminology consists of Low Level Terms (LLTs), Preferred Teams (PTs), High Level Terms (HLTs),

b	i	Type of AE	Group 1 N_1=148	Group 2 N_2=132	Group Diff	2-sided P-value	Posterior $\theta_{bi}>0$	Probability $\theta_{bi}=0$
1	1	Astenia/fatigue	57	40	8.40%	0.167	0.211	0.762
1	2	Fever	34	26	3.30%	0.561	0.122	0.827
1	3	Infection, fungal	2	0	1.40%	0.5	0.101	0.796
1	4	Infection, viral	3	1	1.20%	0.625	0.1	0.813
1	5	Malaise	27	20	3.00%	0.525	0.116	0.826
3	1	Anorexia	7	2	3.20%	0.179	0.117	0.821
3	2	Cendisiasis, oral	2	0	1.40%	0.5	0.083	0.835
3	3	Constipation	2	0	1.40%	0.5	0.101	0.812
3	4	Diarrhea	24	10	8.60%	0.029*	0.231	0.743
3	5	Gastroenteritis	3	1	1.20%	0.625	0.093	0.823
3	6	Nausea	2	7	-3.90%	0.089*	0.05	0.805
3	7	Vomiting	19	19	-1.60%	0.73	0.076	0.849
5	1	Lymphadenopathy	3	2	0.50%	1	0.136	0.717
6	1	Dehydration	0	2	-1.50%	0.221	0.087	0.666
8	1	Crying	2	0	1.40%	0.5	0.185	0.655
8	2	Insomnia	2	2	-0.10%	1	0.153	0.661
8	3	Irritability	75	43	18.10%	0.003*	0.78	0.214
9	1	Bronchitis	4	1	1.90%	0.375	0.059	0.9
9	2	Congestion, nasal	4	2	1.20%	0.375	0.058	0.901
9	3	Congestion, respiratory	1	2	-0.80%	0.603	0.04	0.896
9	4	Cough	13	8	2.70%	0.497	0.062	0.906
9	5	Infection, upper respiratory	28	20	3.70%	0.431	0.083	0.897
9	6	Laryngotracheobronchitis	2	1	0.60%	1	0.047	0.898
9	7	Pharyngitis	13	8	2.70%	0.497	0.061	0.906
9	8	Rhinorrhea	15	14	-0.50%	1	0.051	0.904
9	9	Sinusitis	3	1	1.20%	0.625	0.051	0.903
9	10	Tonsillitis	2	1	0.60%	1	0.042	0.905
9	11	Wheezing	3	1	1.20%	0.625	0.05	0.907
10	1	Bite/sting	4	0	2.70%	0.125	0.087	0.859
10	2	Eczenma	2	0	1.40%	0.5	0.07	0.86
10	3	Pruritis	2	1	0.50%	1	0.062	0.868
10	4	Rash	13	3	6.50%	0.021*	0.19	0.784
10	5	Rash, diaper	6	2	2.60%	0.288	0.099	0.852
10	6	Rash, measles/rubella-like	8	1	4.60%	0.039*	0.126	0.836
10	7	Rash, varicella-like	4	2	1.20%	0.687	0.076	0.862
10	8	Urticaria	0	2	-1.50%	0.221	0.048	0.852
10	9	Viral exanthema	1	2	-0.80%	0.603	0.055	0.855
11	1	Conjunctivitis	0	2	-1.50%	0.221	0.079	0.721
11	2	Otitis media	18	14	1.60%	0.711	0.102	0.757
11	3	Otorrhea	2	1	0.60%	1	0.121	0.749

Table 1: Fisher's 2-sided P-values (with asterisks if <0:1) and posterior probabilities under the Bayesian 3-level hierarchical mixture model.

high level group teams (HLGTs), and System Organ Classes (SOCs). The LLTs constitute the lowest level of terminology and each LLT is linked to one PT. In addition to facilitating data entry and promoting consistency by decreasing subjective choices, the LLTs can also be used for data retrieval without ambiguity because they are more specific than the PTs. A PT must have at least one LLT linked to it, must be linked to at least one SOC, and must have a primary SOC under which the PT appears in data outputs. It is a distinct descriptor for symptom, sign, disease, diagnosis, therapeutic indication, surgical or medical procedure, and medical, social or family history characteristic. As subordinates of HLTs, PTs are linked to HLTs by anatomy, pathology, physiology, etiology or function. Each HLT must be linked to at least one SOC through one of HLGTs, which group HLTs to aid data retrieval at a broader concept.

Gould [6] proposed a three-tier system to categorize adverse events in clinical safety data. Tier 1 is associated with specific hypotheses that are defined by the clinical development team as an adverse event of special interest. Tier 2 is the large set of adverse events encountered as part of the systematic collection and reporting of safety data. The MMRV data summarized above is an example of Tier 2 adverse events. Tier 3 includes the rare spontaneous reports of serious events that require further clinical and epidemiological evaluation. The 40 adverse events from the MMRV trial tabulated in Table 1 are all Tier 2 events. An adverse event can belong to both Tier 1 and Tier 3, and an example is intussusception, which is the telescoping or prolapse of one portion of the bowel into an immediately adjacent segment. Intussusception is an uncommon illness with a background incidence of 18 to 56 cases per 100,000 infant years during the first year of life in the US. In 1998, a tetravalent rhesus-human Reassortant Rotavirus Vaccine (RRV-TV; RotaShield, Wyeth Laboratories) was licensed and recommended by the Advisory Committee for Immunization Practices (ACIP) for routine immunization of infants in the United States. A slight increase in intussusception was observed in the prelicensure studies but did not reach a level of concern. However, post-marketing surveillance studies Murphy et al., [7] showed a temporal association between RRV-TV and intestinal intussusception. As a result of this finding in post-marketing surveillance studies, the RRV-TV vaccine was voluntarily withdrawn from the market in October, 1999 and two weeks later the ACIP rescinded its recommendation for universal vaccination. At the time the intussusception issues arose around the RRV-TV, clinical development of RotaTeq, a pentavalent human-bovine PRV developed by Merck was in Phase II trials. The PRV clinical development program was immediately expanded to include the Rotavirus Efficacy and Safety Trial (REST), which was undertaken to specifically address the safety question on the association between vaccination with the candidate PRV and intussusception. REST was a placebo-controlled study including approximately 70,000 subjects, making it one of the largest clinical trials ever conducted pre-licensure. The clinical importance of REST is discussed in a recent paper by Rosenblatt [8] that highlights the importance and complexity of safety evaluation in clinical development programs for novel drugs and vaccines. Intussusception was considered Tier 3 because it is serious but uncommon in its natural history. Too few cases of intussusception were observed in the original pre-licensure trials of the RRV-TV vaccine to reach a conclusion that could alter the benefit-risk trade-off of an important new vaccine. The association with rotavirus vaccines was established subsequently in post-marketing studies that led to the treatment of intussusception as a Tier 1 adverse event for the subsequent vaccine PRV, for which studies were designed specifically to address the issue prospectively in hypothesis-driven clinical trials. The focus of research on multiplicity issues in the analysis of clinical safety data is related to Tier 2 adverse events, for which the clinical trial data for these are typically summarized by using risk differences, risk ratios, or odds ratios.

False discovery rate and DFDR control

Table 1 summarizes the adverse event data from the MMRV trial by tabulating counts of infants with the specific adverse event type (PT, labelled by i) for body system (SOC, labelled by b), and the between-group risk difference (in %). It also gives a 2-sided P-value computed using Fisher's exact test for each i within body system b. Fisher's exact test is computed from the 2×2 contingency table consisting of the counts n_1, n_2 for the two groups in the first row of the table, and N_1-n_1, N_2-n_2 in the second row of the table. Table 1 shows five (b, i) pairs with one-sided P-value<0.05 (equivalent to two-sided P-value<0.1). Since there are forty (b, i) pairs in Table 1, adjustments have to be made for testing multiple (rather than individual) hypotheses. The ICH E-9 guideline (International Conference on Harmonization or ICH) of technical requirements for regulations of pharmaceuticals for human use [9] discusses this issue and recommends descriptive statistical methods supplemented by individual confidence intervals. It points out that if hypothesis tests are used, statistical adjustments of the type I error for multiplicity may not be appropriate because the type II error is usually of greater concern, and individual P-values may be useful as a flagging device applied to a large number of safety variables to highlight differences worthy of further attention. Hence, the challenge lies in a proper balance between no adjustment and too much adjustment for multiplicity. This has led Mehrotra and Heyse [5] to control the False Discovery Rate (FDR) rather than the more stringent Family-Wise Error Rate (FWER) and to develop a double FDR procedure that further trims down the number of null hypotheses using the body system context. Let $\{H_i, i=1, \bullet\bullet\bullet, m\}$ denote a family of null hypotheses.

In the current setting of adverse event types in a clinical trial, true null hypotheses are those associated with adverse event types for which the incidence is the same between the treatment and control groups. The Family-Wise Error Rate (FWER) is defined as the probability that some true null hypothesis is rejected. Noting that FWER control may be too stringent for many applications, Benjamini and Hochberg [10] propose to control instead the false discovery rate E (V/R), which is the expected proportion of rejected hypotheses that are incorrectly (Table 1).

Table 1 Fisher's 2-sided P-values (with asterisks if <0.1) and posterior probabilities under the Bayesian 3-level hierarchical mixture model. Rejected and in which R is the number of rejected null hypotheses and V is the number of incorrectly rejected Hi. When no hypotheses are rejected (i.e., R=0), the rate (abbreviated by FDR) is defined to be 0. Earlier Soric [11] called rejected hypotheses "statistical discoveries". Since V is the number of false positives, FWER control provides assurance that P (V ≥ 1) does not exceed a prescribed rate α, whereas FDR controls the expected pro-portion of discoveries which are actually false. Note that FWER=P (V ≥ 1) ≤ E (V/R)=FDR. Associated with the m hypotheses in $H_1, H_2, \bullet\bullet\bullet, H_m$ are corresponding unadjusted P-values $P_1, P_2, \bullet\bullet\bullet, P_m$. Let $P_{(1)} \le P_{(2)} \le \bullet\bullet\bullet \le P_{(m)}$ be the ordered P-values, with $H_{(i)}$ corresponding to the hypothesis aligned with $P_{(i)}$. Benjamini and Hochberg have shown that FDR can be controlled at a prespecified rate α by rejecting $H_{(1)}, H_{(2)}, \bullet\bullet\bullet, H_{(J)}$, where J=max{i : $P_{(i)} \le (i/m)\alpha$}, if the P_i are independent. When the above set is empty, no hypotheses are rejected; on the other hand, all hypotheses are rejected if J=m. In comparison with the step-down FWER control procedure that compares $P_{(i)}$ to α/(m+1−i), the FDR procedure compares $P_{(i)}$ to α(i/m). For i=1 and i=m, i/m is equal to 1/(m+1−i), but otherwise i/m is larger,

hence the FDR control procedure should have greater power than the FWER control procedure in detecting the true positives. Mehrotra and Heyse [5] propose to implement the Benjamini-Hochberg procedure by using the adjusted P-values.

$$\tilde{P}_{(m)} = P_{(m)},\ \tilde{P}_{(j)} = \min\{\tilde{P}_{(j+1)}, (\frac{m}{j}) p_{(j)}\}\ for\ j \le m-1$$

Rejecting $H_{(j)}$ if ~ $P_{(j)} \le \alpha$. They also propose a two-stage procedure, called DFDR (double FDR) for aging Tier 2 adverse experiences that are grouped by body systems. The _first stage uses $\tilde{p}_b^* = min(Pb,1,\ldots,\ Pb,mb)$ as the P-value of the bth body system, with mb adverse event types, for b=1; :::;B. These P-values are used to test the null hypothesis $H^{(b)}$ that treatment and control have no differences in the mb adverse event types. They are adjusted for multiplicity (for 1 ≤ b ≤ B), leading to the adjusted P-values $\tilde{p}_3^{(8)} = 0:075$ and the group-level rejection criterion for rejecting $H^{(b)}$ if $\tilde{p}_b^* \le \alpha_2$. The second stage of DFDR applies the Benjamini-Hochberg procedure to the reduced set of null hypothesis $\{H_i^{(b)} : H^{(b)}$ is rejected and $1 \le i \le m_b\}$, leading to adjusted P-value $\tilde{p}_b^*$ and the final rejection criterion for rejecting $H_i^{(b)} \in H$ $\tilde{p}_b^* \le \alpha_2$. Mehrotra and Heyse (2004) propose to choose α_1 and α_2 by bootstrap resampling so that E_{H0} (V/R) ≤ α, where H_0 denotes the intersection null hypothesis $\cap_{b=1}^{B} H^{(b)}$. Instead of a two-dimensional search, they fix $\alpha_{1=}\alpha_2$ or $\alpha_{1=}\alpha_2/2$ and carry out a grid search over $\alpha_2 \le \alpha$.

To illustrate how this two-stage procedure works for the adverse event data in Table 1 from the MMRV combination vaccine safety trial, Table 2 tabulates the unadjusted P-values $\tilde{p}_b^*$ (2-sided, Fisher's exact test) and the corresponding adjusted P-values $\tilde{p}_b^*$. The body system 8 is the only one rejected by the first stage of the DFDR procedure $\tilde{p}_b^* < 0:1$ for b=8). There are 3 adverse event types within b=8: Irritability ($\tilde{p}_3^{(8)} = 0:075$) that is rejected, Crying ($\tilde{p}_2^{(8)} = 1:00$) and Insomnia ($\tilde{p}_2^{(8)} = 1:00$) that are not rejected by the final rejection criterion.

Bayesian approach via a three-level hierarchical mixture model

The last two columns of Table 1 give the results of the posterior probabilities that $\theta_{bi>}0$ and $\theta_{bi=}0$, respectively, under the Bayesian hierarchical mixture model proposed by Berry and Berry [12], where θ_{bi} is the logarithm of the odds ratio of the adverse event probability for treatment (Group 2) to that for control (Group 1):

$$\grave{e}_{bi}=\log\left(\frac{pbi,2}{(1\text{-}pbi,2)}\right)\text{-}\log\left(\frac{pbi,1}{(1\text{-}pbi,1)}\right)$$

where p_{bi}, 1 and p_{bi}, 2 are the adverse event probabilities for Group 1 and Group 2. Note that the column "Group Diff" in Table 1 is the sample estimate of $p_{bi,2}$-$p_{bi,1}$ (Table 2).

The last two columns of Table 1 do not sum up to 1 because there is positive, albeit small, posterior probability that $\theta_{bi<}0$ in the Bayesian model. The first level of the Bayesian hierarchical mixture model assumes that θ_{bi} is 0 with probability πb and is normally distributed with probability 1-πb. The second and third levels of the hierarchical specification gives the prior distributions of πb and of the mean and variance of the normally distributed component of the mixture model at the first level. Berry and Berry [12] point out that their Bayesian specification attempts to model "the existing structure and the available information" among types of Adverse Events (AEs) "explicitly depending on their body systems," thus "borrowing information across types of AEs." Hence, "this is different from conclusions of more traditional multiple comparison methods in which only the number of types of AEs under consideration matters," as in the FDR and DFDR control methods. The Bayesian analysis shows that "the posterior probability that the event rate on treatment is greater than on control is small to moderate (less than 50%) for 39 of the 40 types of AEs," and that there is only one type of AE (irritability in body system 8) with a high value (0.78) for the posterior probability of $\theta_{bi>}0$. This AE type also has the smallest P-value (0.003) for Fisher's exact tests in the individual comparisons shown in Table 1.

A Bayesian screening/classification method

Gould [13] says that "although rejecting a null hypothesis of no treatment effect with suitable adjustment for multiplicity on the basis of predefined measurement in a well-designed- and-executed trial justifies a conclusion that the treatment is effective," this argument does not apply to safety, particularly with respect to Tier 2 adverse events, because "testing hypotheses about treatment group differences in adverse event incidence when the adverse events have not been identified in the study protocol amounts to using observed data to test hypotheses that are generated by the same data." He advocates a Bayesian screening approach that "provides a direct assessment of the likelihood of no material drug-event association and quantifies the strength of the observed association" for the Tier 2 AEs of the control and treatment groups. The screening method proposed is basically a Bayesian classification rule of the form $\theta_{bi} \le \theta^*$ for classifying the observed AE as safe, and flagging safety concerns if $\theta_{bi} > \theta^*$, where θ^* is either "clinically meaningful" to the investigators and regulators or can be determined from the data to yield good diagnostic properties of the classifier. Gould uses another Bayesian mixture model for which posterior probabilities are much easier to compute than Berry and Berry's three-level hierarchical model. Specifically, he assumes that p_{bi}, 2 is equal to p_{bi}, 1 with probability π and has a Beta distribution that is independent of the Beta distribution for p_{bi}, 1 with probability 1-π, and that π also has a Beta distribution. The parameters of the Beta prior distributions are determined from the data so as to strike a good balance between sensitivity and specificity of the classifier.

Body System	# AEs	Representative AE type	Group 1 N_1=148	Group 2 N_2=132	Unadjusted P-value	Adjusted P-value
1	5	Asthenia/fatigue	57	40	0.1673	0.6248
3	7	Diarrhea	24	10	0.0289	0.2026
5	1	Lymphadenopathy	3	2	1	1
6	1	Dehydration	0	2	0.2214	0.2214
8	3	Irritability	75	43	0.0025	0.0075*
9	11	Bronchitis	4	1	0.3746	0.9447
10	9	Rash	13	3	0.0209	0.1745
11	3	Conjunctivitis	0	2	0.2214	0.6641

Table 2: Smallest adjusted P-value from each of the 8 body systems.

Discussion and Conclusion

The past fifteen years witnessed a greatly increased focus on the safety evaluation of medical products in the pharmaceutical and biotechnology industries. Safety data are routinely collected throughout preclinical *in vitro* and *in vivo* experiments (e.g., living cells and animal models), clinical development (e.g., randomized clinical trials) and post-approval studies and monitoring. Whereas most clinical trials are designed to investigate the hypothesized efficacy of a compound, safety outcomes, on the other hand, are often not defined a priori. This brings forth a number of challenges to statisticians and biomedical data scientists on how to best analyze the high-dimensional safety data, in order to detect safety signals promptly and also to reduce the rates of false signals and false non-signals. This commentary reviews some important developments to address these challenges for the analysis of adverse events data from pre-licensure clinical trials and post-marketing phase IV trials. The developments have their roots in contemporary advances in statistical methodology in the big data era, ranging from diverse areas such as FDR control in simultaneous testing of a large number of null hypotheses, Bayesian hierarchical and multi-level models, screening and classification. An overarching approach that can potentially integrate these methods is suggested by the seminal works of Efron et al. [14]; Efron [15-17] on empirical Bayes/compound decision methods and local false discovery rates for the analysis of microarray gene expression data and large-scale simultaneous testing. We are working toward such an approach to clinical safety data evaluation which strikes an optimal balance between sensitivity and specificity.

Before marketing authorization, a medical product is typically investigated thoroughly for safety and efficacy through clinical trials with hundreds or thousands of somewhat homogeneous subjects (sampled from a population with pre-defined inclusion and exclusion criteria) for a relatively short period of time (e.g., 2 years) with clearly specified route of administration. The number of subjects encompassed in such a trial is commonly determined by demonstrating efficacy and rare adverse events may be unobservable. For instance, suppose that the occurrence of an adverse event follows a Poisson distribution. Then the minimum number of subjects (or observational time in person-years) needed in order to observe at least 1 reported case of a target adverse event with an incidence rate at 0.1% with 95% confidence is approximately 2996; the number of subjects (or person-years) goes up to at least 4744 in order to observe at least two reported cases of the target adverse event with the same incidence rate. In addition to relatively smaller sample size, there are usually quite strict inclusion and exclusion criteria for subject enrollment in clinical trials; hence co-morbidity and/or drug-drug interactions may not be discovered during clinical trials [18]. Because of these limitations of clinical trials, safety evaluation of medical products is usually carried out after the pre-licensure and post-marketing clinical trials through the whole life of a product. When post-marketing safety data come from non-experimental sources, as in spontaneous reports of adverse events rather than randomized trials, there may be confounding covariates that cause the adverse events and adjustments have to be made for causality analysis. This poses important methodological challenges that are beyond the scope of the present commentary on sensitivity versus specificity in testing multiple safety hypotheses, or in classifying (screening) the adverse events from the clinical trials data as safe or unsafe outcomes. Again contemporary developments in statistical methods and in pharmacoepidemiology provide many important techniques that can potentially be integrated to address the challenges of using these safety databases for pharmacovigilance and syndromic surveillance. Propensity scores, graphical models, instrumental variables, and inverse probability weighting are a partial list of the statistical methods. A corresponding list for pharmacoepidemiology includes assessment of medication adherence and medication errors (or of device misuse or malfunctioning leading to device- related adverse experiences for medical devices), reporting ratios and disproportionality analysis, case-control approach and self-controlled case series.

References

1. Xia HA, Ma H, Carlin BP (2011) Bayesian hierarchal modeling for detecting safety signals in clinical trials. J Biopharm Stat 21: 1006-1029.
2. Chuang-Stein C, Xia HA (2013) The practice of pre-marketing safety assessment in drug development. J Biopharm Stat 23: 2-25.
3. Crowe BJ, Xia HA, Berlin JA, Watson DJ, Shi H, et al. (2009) Recommendations for the 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 6: 430-440.
4. Kaplan KJ, Rusche SA, Lakkis HD, Bottenfield G, Guerra FA, et al. (2002) Post-licensure comparative study of unusual high-pitched crying following Comvax and Placebo versus Pedvax HIB and Recombivax HB in healthy infants. Vaccine 21: 181-187.
5. Mehrotra DV, Heyse JF (2004) Use of the false discovery rate for evaluating clinical safety data. Stat Methods Med Res 13: 227-238.
6. Gould AL (2002) Drug safety evaluation in and after clinical trials. Deming Conference, Atlantic City, NJ, USA.
7. Murphy TV, Gargiullo PM, Massoudi MS, Nelson D, Jumaan A, et al. (2001) Intussusception among infants given an oral rotavirus vaccine. N Engl J Med 344: 564-572.
8. Rosenblatt M (2017) The changing face of clinical trials: The large pharmaceutical company perspective. N Engl J Med 376: 52-60.
9. International Conference on Harmonization (1998) Technical requirements for regula- tions of pharmaceuticals for human use. Ich triparlite guideline e-9 document. Statistical Principles for Clinical Trials.
10. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57: 289-300.
11. Soric B (1989) Statistical discoveries and effect size estimation. J Am Stat Assoc 84: 608-610.
12. Berry SM, Berry DA (2004) Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics 60: 418-426.
13. Gould AL (2008) Detecting potential safety issues in clinical trials by Bayesian screening. Biom J 50: 837-851.
14. Efron B, Tibshirani R, Storey J, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96: 1151-1160.
15. Efron B (2003) Robbins, empirical Bayes and microarrays. Ann Stat 31: 366-378.
16. Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99: 96-104.
17. Efron B (2007) Size, power and false discovery rate. Ann Stat 35: 1351-1377.
18. Trontell A (2004) Expecting the unexpected drug safety, pharmacovigilance, and the prepared mind. N Engl J Med 351: 1385-1387.