Yonsei Med J 48(6):895 - 900, 2007
DOI 10.3349/ymj.2007.48.6.895

Review Article

Rethinking Statistical Approaches to Evaluating Drug Safety

Jen-pei Liu[1,2]

[1]Statistical Education Center; Consulting Center for Statistics and Bioinformatics; Division of Biometry, Graduate Institute of Agronomy, National Taiwan University, Taipei, Taiwan; and [2]Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan.

Purpose: The current methods used to evaluate the efficacy of drug products are inadequate. We propose a non-inferiority approach to prove the safety of drugs. **Materials and Methods:** Traditional hypotheses for the evaluation of the safety of drugs are based on proof of hazard, which have proven to be inadequate. Therefore, based on the concept of proof of safety, the non-inferiority hypothesis is employed to prove that the risk of new drugs does not exceed a pre-specified allowable safety margin, hence proving that a drug has no excessive risk. The results from papers published on Vioxx® and Avandia® are used to illustrate the difference between the traditional approach for proof of hazard and the non-inferiority approach for proof of safety. **Results:** The *p*-values from traditional hypotheses were greater than 0.05, and failed to demonstrate that Vioxx® and Avandia® are of cardiovascular hazard. However, these results cannot prove that both Vioxx® and Avandia® are of no cardiovascular risk. On the other hand, the non-inferiority approach can prove that they are of excessive cardiovascular risk. **Conclusion:** The non-inferiority approach is appropriate to prove the safety of drugs.

Key Words: Effectiveness, safety, no excessive risk, non-inferiority approach

INTRODUCTION

After more than 100 deaths caused by the Elixir Sulfanilamide disaster in 1938, the US Congress passed the Federal Food, Drug, and Cosmetic Act (FD&C Act) which, for the first time in US history, required pharmaceutical companies to submit full reports of investigations regarding the safety of new drugs. However, it was not until 1962, after the passage of the Leaver-Harris Amendment of the FD&C Act, that the US Food and Drug Administration (US FDA) was authorized to require evidence of efficacy for approval of new drugs. Consequently, for the approval of a new drug, the US FDA requires adequate and controlled clinical trials be conducted on humans to demonstrate their effectiveness and safety. The safety of a drug ought to be the primary focus and should come before its efficacy.

However, there has recently been an alarmingly increasing trend of safety issues of drugs after their approval. For instance, the most notorious example is the withdrawal of Vioxx® (rofecoxib) in 2004 after its excessive cardiovascular risk was not only confirmed by a retrospective nested case-control study[1] but also by prospective clinical trials.[2,3] The most recent safety saga is also the cardiovascular risk associated with Avandia® (rosiglitazone), one of the most widely used thiazolidinediones that are agonists for peroxisome-proliferator-activated receptor ɣ (PPAR-ɣ. Nissen and Wolski[4] reported a meta-analysis that showed that the risk of myocardial infarction is 43% higher than that of the control group. These safety issues reflect that the current evaluation processes are not adequate for the assessment of the safety of drugs for approval. Consequently, an US congressional hearing and an US FDA joint advisory committee meeting were called on June 6 and July 30, 2007, respectively, to review the issues surrounding drug safety.[5]

As mentioned before, safety comes before efficacy. However, at least in design, conduct and analysis of the adequate and well-controlled

Received October 23, 2007

The views expressed in this article are the opinions of the author and do not necessarily represent the views of National Taiwan University and the National Health Research Institutes, Taiwan.

Reprint address: requests to Dr. Jen-pei Liu, Division of Biometry, Department of Agronomy, National Taiwan University, 1, Section 4, Roosevelt Road, Taipei, Taiwan, Tel: 886-2-3366-4791, Fax: 886-2-3366-4791, E-mail: jpliu@ntu.edu.tw

clinical trials, this may not be totally true for pharmaceutical companies to develop new drugs and for regulatory agencies to approve them. The current paradigm for the approval of a new drug is two-fold. First, the drug must be proven to be efficacious. Second, it must be verified whether there is any excessive safety risk even though the objective is to prove that the drug is safe. As a result, the selection of study design, endpoints, statistical methods, and sample size are to maximize the probability of proving the effectiveness of the drug. On the other hand, most clinical trials conducted during the development of the drug do not select the optimal design with a sufficient number of patients and correct endpoints to prove that the drug is safe. In addition, most of the analyses of safety data are descriptive in nature and no inferences are made. Consequently, these trials are neither adequate nor well controlled for the evaluation of safety.

One of the most critical but often neglected components during the review process for drug approval is the statistical evaluation of the evidence of safety. To prove the effectiveness of a drug, the approach is to adopt the traditional hypothesis of equality. In other words, the effectiveness of a drug is proven by rejecting the null hypothesis of equal efficacy between the test drug and control to prove that the alternative hypothesis of a superior efficacy of the test drug is true. This approach is adequate for proof of efficacy of drug products. However, the same approach is not appropriate for the evaluation of safety because it is for proof of hazard or excessive risk of drugs.[6] Therefore, failure to reject the null hypothesis of no excessive risk cannot prove that a drug is safe.[7] To prove that a drug is safe, based on the concept of risk management, we suggest that the non-inferiority hypothesis with consideration of managing the magnitude of the safety risk is more appropriate than the traditional hypothesis.

MATERIALS AND METHODS

Current approaches

For the sake of illustration, we considered a situation in which a clinical trial with a randomized two-group parallel design was conducted to compare the efficacy and safety of a test drug with a concurrent control group. Let R denote the risk ratio (relative risk) or odds ratio of the test drug compared to the control with respect to a pre-defined adverse event (AE) such as confirmed cardiovascular events. Currently, the inference of safety evaluation is based on the following traditional hypothesis:

$$H_o: R \leq 1 \text{ vs. } H_a: R > 1. \quad (1)$$

The null hypothesis H_o in equation (1) states that the risk of the test drug is smaller than that of the control. On the other hand, the alternative hypothesis H_a in equation (1) states that the risk of the test drug is greater than that of the control. Therefore, the objective of this formulation of the hypothesis for evaluation of safety is to verify whether the test drug is of excessive risk with respect to a pre-defined AE compared to the control. This approach is referred to as the proof-of-hazard approach.[6] When the null hypothesis H_o in equation (1) is rejected at the α significance level, it can then be concluded that the test drug is of excessive risk. However, when the null hypothesis in equation (1) is not rejected, the only conclusion that can be reached is that the data can not provide sufficient evidence to doubt the validity of the null hypothesis.[7] In other words, failure to reject the null hypothesis does not necessarily prove the null hypothesis of no excessive risk in equation (1) and cannot conclude that the test drug is safe.

The decision based on the hypothesis in equation (1) is either to reject the null hypothesis or fail to reject the null hypothesis. Consequently, another drawback of the formulation of the hypothesis in equation (1) is its qualitative nature because it does not take the magnitude of the risk into consideration. The other two disadvantages of the hypotheses in equation (1) are well known in the area of bioequivalence.[8] If a study was poorly conducted, a non-significant result may be due to a larger variability associated with the safety data although that the risk of the test drug is in fact greater than that of the control. Furthermore, sample sizes for most drug trials currently are powered only to demonstrate the

efficacy of the test drug and may not provide sufficient power for evaluation of safety. A non-significant result regarding the safety does not prove that the test drug is safe. In summary, the traditional hypotheses in equation (1) cannot prove that the drug is safe, and fails to take into account the magnitude of the risk.

Non-inferiority approach

The objective of the evaluation of safety data for approval of a new drug is to prove that the drug is safe. In other words, one should prove that the test drug poses no excessive risk in the targeted patient population compared to the control group. This concept is referred to as the proof of safety.[6] Therefore, we suggest that the inferential assessment of the safety data is formulated as the following non-inferiority hypothesis[9]

$$H_o: R > 1 + \Delta_0 \text{ vs. } H_a: R \leq 1+\Delta_0, \qquad (2)$$

where $\Delta_0 > 0$ is some pre-specified safety margin based on the relative risk.

The alternative hypothesis H_a in equation (2) is to prove by a clinically allowable and inconsequential margin that the risk associated with a pre-defined adverse event of the test drug is not greater than that of the control group. Therefore, the formulation of the non-inferiority hypotheses manages not only the magnitude of the safety risk within an acceptable margin but also correctly expresses the hypothesis of no excessive risk as the alternative hypothesis H_a in equation (2). It follows that rejecting the null hypothesis H_o in equation (2) proves that the test drug is of no excessive risk.

Statistical methods for testing the non-inferiority hypotheses are available and can directly be applied to the evaluation of the safety of the test drug.[10-15] However, a more informative way to test the non-inferiority hypotheses in equation (2) is to construct a $(1 - 2\alpha)100\%$ confidence interval (CI) for the risk ratio. If the upper limit of the $(1 - 2\alpha)$ 100% CI for the risk ratio is less than the pre-specified allowable safety margin in the targeted patient population, then at the α significance level, the test drug can be concluded of no excessive risk as compared to the control group. The confidence interval approach is preferable because it can test not only the non-inferiority hypotheses in equation (2) but also provides a quantitative range of the risk ratio with $(1 - 2\alpha)100\%$ confidence.

We applied both the traditional and non-inferiority hypotheses to the results of the studies from papers published on Vioxx® and Avandia®. For the purpose of illustration of the difference between the concepts of proof of hazard and proof of safety, α was selected as 2.5% and Δ_0 in equation (2) was chosen to be 0.5 for which a 50% increase of safety risk of the test drug over the control is clinically allowed.

RESULTS

Graham et al.[1] reported the results of a nested case control study on cardiovascular risk of COX-2 inhibitors. This study is based on a claim database of a national integrated managed care organization in conjunction with the mortality status and cause of death from the California Department of Health and Center for Health Statistics. Table 1 shows the odds ratios of acute myocardial infarction with the use of selected NSAIDs compared to remote use of a NSAID. Table 1 reveals that the upper limits of the 95% CI for odds ratios of acute myocardial infarction for celecoxib and ibuprofen are less than 1.50. It follows that, with respect to the risk of acute myocardial infarction, celecoxib and ibuprofen pose no excessive risk compared to remote users of NSAIDs. For naproxen, even though the *p*-value for the hypothesis in Equation (1) is 0.05, it can be concluded that naproxen is of no excessive risk of acute myocardial infarction compared to remote users of NSAIDs, because the upper limit of 95% CI for odds ratios of acute myocardial infarction is less than the upper safety margin of 1.5. On the other hand, since their upper limits of all 95% CIs for odds ratios are larger than the safety margin of 1.5, it cannot be concluded that rofecoxib is of no excessive risk of acute myocardial infarction over the remote users of NSAIDs even though some *p*-values based on the hypothesis in equation (1) is greater than 0.05.

Bresalier et al.[3] reported that the risk of cardiovascular event associated with rofecoxib in the APPROVe trial, which was a chemoprevention

Table 1. Odds Ratios of Acute Myocardial Infarction with Use of Selected NSAIDs Compared with Remote Use of a NSAID

NSAID	Odds ratio	95% C.I.	Adjusted p value based on hypothesis in Eq. (1)
Remote use	1		
Celecoxib	0.84	(0.67, 1.04)	0.12
Ibuprofen	1.06	(0.96, 1.17)	0.27
Naproxen	1.14	(1.00, 1.30)	0.05
Rofecoxib (all doses)	1.34	(0.98, 1.82)	0.066
Rofecoxib			
<= 25 mg/day	1.23	(0.89, 1.71)	0.21
> 25 mg/day	3.00	(1.09, 8.31)	0.03

Adapted from Graham, et al.[1]

Table 2. Relative Risks of Confirmed Serious Thrombotic Events of Rofecoxib Compared with Placebo in APPROVe Trial

Adverse event	Relative risk	95% C.I.
Overall all	1.92	(1.19, 3.11)
Month 0 - 18	1.18	(0.64, 2.15)
Month 19 - 36	4.45	(1.77, 13.32)

Adapted from Bresalier, et al.[3]

trial involving 2,586 subjects with colorectoral adenoma. Table 2 provides the relative risks of confirmed serious thrombotic events of rofecoxib compared to the placebo group. Table 2 shows that although the 95% confidence interval of relative risk includes 1 between month 0 and month 18, its upper limit is 2.15, which is greater than the safety margin of 1.5. Therefore, contrary to the claim made by the paper, rofecoxib cannot be concluded to be of no excessive risk with respect to serious thrombotic events.

Table 3 presents partial results by Nissen and Wolski[4] who performed a meta-analysis on a total of 26,000 patients for the effects of Avandia® on the risk of myocardial infarction and death from cardiovascular causes. As seen in Table 3, the upper limits of all 95% CIs for odds ratios are greater than 1.5. Despite the fact that some of the 95% CIs for odds ratios include 1, it cannot be concluded that Avandia® possesses no excessive risk of myocardial infarction and death from cardiovascular causes. In response to the meta-analysis by Nissen and Wolski,[4] Home et al.[16] reported an interim analysis of the RECORD trial of a total of 4,000 patients with a planned median follow-up of 6 years for the effect of Avandia® on cardiovascular outcomes. The hazard ratios of death from cardiovascular causes, acute myocardial infarction, or congestive heart failure are given in Table 4. With respect to death from various causes, the upper limits of all 95% CIs for hazard ratios are less than 1.5. Therefore, from the data accumulated at the cut-off date of March 30, 2007, with a mean follow-up of 3.75 years, Avandia® can be concluded to be of no excessive risk of death. On the other hand, however, the upper limits of the 95% CIs of hazard ratios with respect to acute myocardial infarction, and congestive heart failure are greater than 1.5. Therefore, Avandia® cannot be claimed to be of no excessive risk of acute myocardial infarction and congestive heart failure.

Table 3. Risk of Myocardial Infarction and Death from Cardiovascular Causes of Avandia®

Study	Odds ratio	95% C.I.	p value based on hypothesis in Eq. (1)
Myocardial infarction			
Small trials combined	1.45	(0.88, 2.39)	0.15
DREAM trial	1.65	(0.74, 3.68)	0.22
ADOPT trial	1.33	(0.80, 2.21)	0.27
Overall	1.43	(1.01, 1.98)	0.03
Death from CV causes			
Small trials combined	2.40	(1.17, 4.91)	0.02
DREAM trial	1.20	(0.52, 2.78)	0.67
ADOPT trial	0.80	(0.17, 3.86)	0.78
Overall	1.64	(0.98, 2.74)	0.06

Adapted from Nissen and Wolski.[4]

Table 4. Hazard Ratios of Death from Cardiovascular Causes of Avandia® in RECORD Trial

Adjudicated events	Hazard ratio	95% C.I.	p value based on hypothesis in Eq. (1)
Death			
From CV causes	0.83	(0.51, 1.36)	0.46
From any cause	0.93	(0.67, 1.27)	0.61
Acute myocardial infarction	1.16	(0.75, 1.81)	0.50
Congested heart failure	2.24	(1.27, 3.97)	0.006
Death from CV causes, myocardial infarction, and stroke	0.97	(0.73, 1.29)	0.83

Adapted from Home, et al.[16]

DISCUSSION

For approval of a new drug, sponsors and regulatory agencies must assure that the drug is safe. However, the continuing saga of emerging safety problems after approval warrants us to rethink and review the current approach to evaluation of the safety data before approval. As formulated by the traditional hypothesis in equation (1), the current paradigm for evaluation of safety is based on the concept of proof of hazard to detect existence of excessive safety risk of the new drug. Therefore, failure to reject the null hypothesis of no excessive risk does not necessarily prove that the new drug is of no excessive safety risk and hence cannot guarantee that it is safe. In addition, the magnitude of safety risk has not been considered in the traditional hypotheses in equation (1). As a result, the current statistical approach to assessing the drug safety is neither appropriate nor adequate.

The notion that a drug is safe implies that the drug is of no excessive safety risk in the targeted patient population. In other words, no excessive safety risk means that the safety risk of the drug can not exceed a clinically inconsequential margin with respect to the control group in the targeted patient population. Therefore, the non-inferiority hypothesis with consideration of the magnitude of the safety risk in equation (2) is more appropriate

to prove that the drug is of no excessive risk. In addition, statistical methods for the non-inferiority hypothesis are available and can directly be applied to verify the safety of the drug.

The selection of safety margins is an extremely important issue in the implementation of the non-inferiority hypothesis for evaluation of safety of drug products. Safety margins depend on the seriousness and consequences of the adverse events, targeted patient population, duration of intended use, magnitude of effectiveness, desirable benefit-risk ratio, and many other factors. Therefore, safety margins should vary depending on different diseases, classes of drugs, and targeted patient population. They should be determined jointly by clinicians, epidemiologists, pharmacists, statisticians, and other personnel involved in drug development as well as approval processes from sponsors, academia, and regulatory agencies. During phase I and II trials, information on safety profiles of new drugs can be obtained. Therefore, for phase III studies, the protocol should specify the primary safety parameters and safety margins in addition to the primary efficacy endpoints. Furthermore, phase III trials should be powered to verify that the drug is of no excessive risk based on the non-inferiority hypothesis. Only in this way, the safety of drugs and patients can be guaranteed.

REFERENCES

1. Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, et al. Risk of acute myocardial infarction and sudden death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 2005;365:475-81.
2. Topol EJ. Failing the public health-rofecoxib, Merck, and FDA. N Engl J Med 2004;351:1707-9.
3. Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 2005;352:1092-102.
4. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007;356:2457-71.
5. Harris G. FDA issues strictest warning on diabetes drugs. New York Times 2007 Jun 6. Available from: http://www.nytimes.com/2007/06/07/health/07drug.htm.
6. Hauschke D, Hothorn LA. Safety assessment of toxicological studies: proof of safety versus proof of hazard. In: Chow SC, Liu JP, editors. Design and analysis of animal studies in pharmaceutical development. New York: Marcel Dekker, Inc.; 1995. p.197-225.
7. Colton T. Statistics in medicine. Boston (MA): Little, Brown and Company; 1974.
8. Chow SC, Liu JP. Design and analysis of bioavailability and bioequivalence studies. 2nd ed. New York (NY): Marcel Dekker, Inc.; 2000.
9. Chow SC, Liu JP. Design and analysis of Clinical Trials. 2nd ed. New York (NY): John Wiley and sons; 2004.
10. Farrington CP, Manning G. Test statistics and sample size formulae for comparing binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990;9:1447-54.
11. Liu JP, Hsueh HM, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired binary data. Stat Med 2003;21:231-45.
12. Hsueh HM, Liu JP, Chen JJ. Unconditional exact tests for equivalence or non-inferiority for paired binary endpoints. Biometrics 2001;57:478-83.
13. Chan ISF. Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. Stat Methods Med Res 2003;12:37-58.
14. Liu JP, Fan HY, Ma MC. Tests for equivalence based on odds ratio for matched-pair design. J Biopharm Stat 2005:15:889-901.
15. Tang NS, Tang ML, Wang SF. Sample size determination for matched-pair equivalence trials using rate ratio. Biostatistics 2007;8:625-31.
16. Home PD, Pocock SJ, Beck-Nielsen H, Gomis R, Hanefeld M, Jones NP, et al. Rosiglitazone evaluated for cardiovascular outcomes-an interim analysis, N Engl J Med 2007;357:28-38.