Chapter 10
Collective Evidence in Drug Evaluation

Qian H. Li

Disclaimer: The views presented in this chapter are the author's own views.

Abstract Multiple doses, endpoints, and tests are used in several clinical studies to establish drug efficacy. Statistical evaluation relies heavily on multiplicity adjustments within one study to control the type I error rate. The use of multiplicity adjustment procedures (MAPs) sometimes leads to conclusions that may not seem logical. As drug efficacy evaluation involves aspects such as assessing efficacy, selecting optimal doses, and labeling claims, incorporating all the aspects under the umbrella of controlling type I error may not be an optimum strategy. Alternatively, a practical approach that uses collective evidence is proposed to evaluate multiple studies, doses, endpoints, and tests. Instead of controlling the type I error, specific types of errors are controlled, such as the error of wrongly approving an ineffective drug and the error of labeling false information. With the collective evidence approach, the need of MAPs in individual studies is debated when multiple studies are available.

10.1 Introduction

Drug efficacy evaluation usually is based on evidence from multiple clinical studies that assess multiple doses using multiple endpoints and tests. The multiplicity issues arising from the clinical studies are classic problems in drug evaluation and have been heavily studied by the regulatory agencies, pharmaceutical and biotech industries, and research institutes (Chuang-Stein et al. [3]; Committee for Proprietary Medical Products (CPMP) by EMEA [4]; Pocock [16]; Proschan and Waclawiw [17]; Shih and Quan [18]). The majority of statistical methods, such as the closed testing procedure

Q. H. Li (✉)
National Institute of Health, National Center for Complementary and Alternative Medicine,
Democracy Blvd., Suite 401, Bethesda, MD 20892-5475, USA
Tel.: 301-594-8018 6707
e-mail: Qian.li@nih.gov

Z. Chen et al. (eds.), *Applied Statistics in Biomedicine and Clinical Trials Design,*
ICSA Book Series in Statistics, DOI 10.1007/978-3-319-12694-4_10

(Marcus et al. [12]), Bonferroni correction, and Hochberg procedure (Hochberg [7]), referred to as multiplicity adjustment procedures (MAPs), have been developed based on the logic that multiplicity, such as multiple doses, endpoints, or tests, increases the chance of detecting a statistically significant result from an ineffective drug. Commonly used analogies of the multiplicity issues have been situations such as betting on a horse race or buying lottery tickets, where increasing the number of horses that one bets on, or the number of tickets that one purchases, would increase chances of winning.

These horse race and lottery analogies have, at times, misled the understanding of multiplicity issues in drug evaluation and disguised rudimentary differences between drug evaluation and such games of chance. In a race, the determination of the winning horse does not depend upon the distance between the winner and the losing horses. That is, the relative speeds of the losing horses do not matter. However, in drug evaluation, the efficacy determination of the drug depends upon the collective performance of individual doses, endpoints, and studies. If only one dose shows a statistically significant treatment effect while none of the other doses show any trend of efficacy, the evidence is less convincing for an efficacious drug than the case where multiple doses show trend of efficacy. Therefore, multiplicity in drug evaluation may not necessarily increase the chance to claim that an ineffective drug is efficacious when information is evaluated collectively.

In this chapter, a practical approach to evaluate evidence collectively is proposed. This method controls the specific types of errors encountered in drug evaluation, such as the error rate of wrongly approving an ineffective drug and the error rate of labeling false information. Moreover, it controls the consistency of evidence. Sect. 10.2 discusses the problems of applying MAPs. Section 10.3 presents the concept of the collective evidence and describes the practical approach. Section 10.4 covers the application of collective evidence in cases of multiple studies, doses, endpoints (including co-primary endpoints and secondary endpoints) and tests. Two cases are discussed in Sect. 10.5 to illustrate the use of collective evidence in understanding the effect of drugs. Throughout this chapter, one-sided p values and one-sided statistical significant levels are used unless otherwise specified.

10.2 Problems of Applying MAPs

Prespecification is vital in the protocol development to ensure careful planning in study design, experiment procedures, endpoint selection, and statistical analysis plans, etc. However, it can be problematic to prespecify decision rules, which are MAPs, in the individual study protocols. The intention of the prespecified decision rule is to reduce the chance of claiming success, yet the selection of the decision rules appears somewhat arbitrary. The same study results may reach different conclusions depending upon the choice of the decision rules. For instance, p values 0.040 and 0.012 were observed for high and low doses, respectively, in a study. If the closed testing procedure using high dose to protect low dose was prespecified, the results

Table 10.1 *P* values of two studies with two doses in each study

	1-sided *p* values	
	High dose	Low dose
Study 1	0.028	0.015
Study 2	0.025	0.013

would not pass the decision rule and would yield an inconclusive conclusion. However, if either the Hochberg procedure or Bonferroni correction was prespecified, the low dose would be considered to be efficacious. Clearly, these distinct conclusions are the result of the prespecified decision rules, which are inflexible and, to a certain extent, arbitrary.

Another problem is that the MAPs may overvalue the isolated effect. To illustrate this, consider a study with three parallel doses and a control arm. If both high and medium doses yielded *p* values of 0.500 and the low dose yielded 0.001, both the Hochberg and Bonferroni procedures would conclude that the low dose was efficacious, despite the fact that there was no sign of efficacy in the other doses. Unless other information supported that this drug had narrow therapeutic window, the evidence would not be considered convincing. Whereas if three doses from high to low yielded *p* values of 0.028, 0.025, and 0.015, respectively, some MAPs would consider such evidence inconclusive. Thus, only looking at the performance of the individual doses rather than the totality evidence may not lead to useful conclusions.

The problem can be more confusing when data from more than one study are available. In fact, two phase 3 studies have been the requirement by the US Food and Drug Administration (FDA) for the purpose of establishing substantial evidence (US FDA [19]; US FDA [20]). Suppose that two phase 3 studies were conducted to support a claim. Also, suppose that two doses, high and low, were included in both studies and a closed testing procedure using high dose to protect low dose was placed in each study. The *p* values of the two doses from both studies were listed in Table 10.1. Following the closed testing procedure, study 1 would be concluded as a "failed" study since it failed to pass the closed testing procedure, whereas study 2 would be considered a successful study. However, the fundamental question of the efficacy of the drug has not been answered.

The application of the MAPs is to protect the type I error. However, the meaning of the type I error is not clear since it covers different types of errors that may occur in various aspects and stages of drug evaluation. Errors can occur when deciding if a drug works, selecting the optimal doses, and labeling drug information with selective endpoints, etc. When deciding if a drug is efficacious, it is necessary to control the error rate of wrongly approving an ineffective drug. When selecting the optimal doses, it is necessary to reduce the error rate of selecting suboptimal doses. When labeling drugs, it is necessary to limit the error rate of providing false information. These different types of errors play different roles in the drug evaluation process and may not necessarily be controlled simultaneously. It is easy to understand that the error of selecting suboptimal doses, or the error of false labeling information would not occur if the drug is concluded to be ineffective. On the other hand, the error made in efficacy decisions should not be impacted by the decision of selecting the optimal doses and the decision of drug labeling. Therefore, it may be less confusing to differentiate the types of errors in drug evaluation and control the different types of error rates separately.

10.3 Concept of Collective Evidence

10.3.1 The Two Types of Logic

In preparation for discussing the alternative approach proposed, two types of logic are considered. Mathematically, the "OR" logic is the union of all events and represented as $E_1 \cup E_2 \cup \ldots \cup E_K$; the "AND" logic is the intersection of all events and formulated as $E_1 \cap E_2 \cap \ldots \cap E_K$, where E_k, $k = 1, 2, \ldots, K$, are events. The K events can be the number of bets that is put down in a horse race for example, or K doses, K endpoints, or K individual studies in drug evaluation. The "OR" logic is the basis for most of the MAPs where success is claimed if one event out of the K events is true. On the contrary, the success definition with the "AND" logic requires that all events are true. The main feature of the collective evidence approach is to include the "AND" logic.

A discussion of Fig. 10.1 illustrates the concept of collective evidence. In Fig. 10.1, the blue area represents the rejection region of Bonferroni correction to control the error rate at the level of 0.025 for two independent p values p_1 and p_2. The Bonferroni correction can be written as $P(p_1 \leq 0.0125 \cup p_2 \leq 0.0125) < 0.025$ or can be written as $P(p_{(1)} \leq 0.0125 \cap p_{(2)} \leq 1.000) < 0.025$ where $p_{(1)}$ and $p_{(2)}$ are ordered p values of p_1 and p_2. Notice that the Bonferroni correction is rewritten using the AND logic, although it can be simplified to $P(p_{(1)} \leq 0.0125) < 0.025$ as the event of a p value less than 1 is always true. Using γ_1, γ_2 to denote the p value cut points for the ordered p values, respectively, the decision rule for the Bonferroni correction can be written as $(\gamma_1, \gamma_2) = (0.0125,1.000)$. This rejection region allows the success claim if one of the p values is 0.0125 or less. It is important to understand that the Bonferroni correction is not the only way of controlling the error rate at the level of 0.025. The green area represents another rejection region that controls error rate at the level of 0.025, that is, $P(p_{(1)} \leq 0.025 \cup p_{(2)} \leq 0.5125) \leq 0.025$. The decision rule is $(\gamma_1, \gamma_2) = (0.025,0.5125)$. This green rejection region supports the success claim if the smaller p value is less than or equal to 0.025 and the larger p value is less than 0.5125. The orange area represents yet another rejection region that controls the same error rate with decision rule $(\gamma_1, \gamma_2) = (0.050,0.275)$. This decision rule covers the rejection region that allows the smaller p value to be 0.050 or less and the larger one has to be 0.275 or less. Notice that both the orange and green rejection regions use the AND logic.

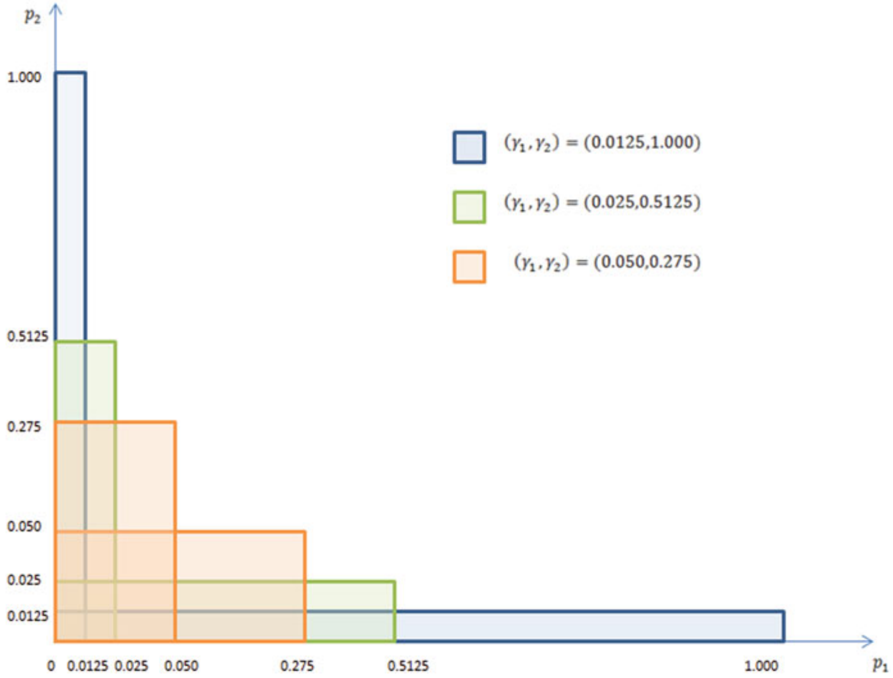

Fig. 10.1 Rejection regions that control the error at the level of 0.025 under different decision rules

10.3.2 The Formulation of the Collective Evidence Approach

The concept of collective evidence was originally proposed by Li and Huque (Li and Huque [10]) for the purpose of evaluating multiple studies and was extended to the evaluation of co-primary endpoints by Li (Li [9]). The concept of collective evidence approach can be described as follows:

1. Similar to a single hypothesis testing scenario, individual null and alternative hypotheses, H_{0k} and H_{Ak}, are used to test each individual event E_k and the test yields p value $p_k, k = 1, 2, \ldots, K$. Null represents no effect, while the alternative is the complement. The K p values are ranked as $p_{(1)} \le p_{(2)} \le \ldots \le p_{(K)}$.
2. Next is to formulate the overall hypothesis to test if a drug works. An overall null hypothesis represents the case that all the individual null hypotheses H_{0k}, $k = 1, 2, \ldots, K$, are true, i.e., an ineffective drug. The corresponding alternative is that at least one of the individual alternatives is true. The overall null and corresponding alternative are denoted as $H_0^{1/K}$ and $H_A^{1/K}$, respectively, and formulated as

$$H_0^{1/K} : \bigcap_{k=1}^{K} H_{0k} \text{ versus } H_A^{1/K} : \bigcup_{k=1}^{K} H_{Ak}$$

The error of wrongly rejecting the overall null is defined as

$$P_{H_0^{1/K}}(p_{(1)} \leq \gamma_1 \cap p_{(2)} \leq \gamma_2 \cap \ldots \cap p_{(K)} \leq \gamma_K),$$

where $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_K$ is a set of the decision rule. The error rate is controlled at the level of $\alpha(H_0^{1/K})$ under $H_0^{1/K}$. When the events are independent studies, $\alpha(H_0^{1/K})$ represents the error rate of wrongly approving an ineffective drug. For two studies, the error rate $\alpha(H_0^{1/2})$ is usually controlled at the level of $0.025^2 = 0.000625$ (Li and Huque [10]). This level of error rate arises from the requirement of two statistically significant studies as the substantial evidence for drug approval. When the events are correlated endpoints within one study, $\alpha(H_0^{1/K})$ represents the error rate of wrongly claiming an ineffective drug to be efficacious. For two co-primary endpoints in one study, the error rate $\alpha(H_0^{1/2})$ is controlled at the level of 0.025 (Li [9]).

To further illustrate the point discussed here, Fig. 10.2 presents two rejection regions for two independent studies in the coordinates p_1 and p_2, representing results of the two studies. Both rejection regions control the error rate at the level of 0.000625. The orange area represents the rejection region for decision rule $(\gamma_1, \gamma_2) = (0.025{,}0.025)$ and the green area represents the rejection region for decision rule $(\gamma_1, \gamma_2) = (0.010{,}0.036)$. Therefore, if two studies yielded p values of (0.010,0.030), this could be considered as convincing evidence for an efficacious drug.

3. In addition, another set of overall hypotheses is formulated to test if all events present efficacy—a reflection of consistency among all events. The overall alternative requires that all events show efficacy. The overall null is therefore that at least one event does not have efficacy. The overall null and alternative are denoted as $H_0^{K/K}$ and $H_A^{K/K}$, respectively, and formulated as

$$H_0^{K/K} : \bigcup_{k=1}^{K} H_{0k} \text{ versus } H_A^{K/K} : \bigcap_{k=1}^{K} H_{Ak}$$

The level of the error should be controlled at the level of $\alpha(H_0^{K/K})$ under the overall null hypothesis $H_0^{K/K}$. It has been shown (Li [9]) that the error rate $\alpha(H_0^{K/K})$ of rejecting the null is

$$P_{H_0^{K/K}}(p_{(1)} \leq \gamma_1 \cap p_{(2)} \leq \gamma_2 \cap \ldots \cap p_{(K)} \leq \gamma_K) \leq \gamma_K,$$

where γ_K is the largest p value cut point of the decision rule $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_K$. For the decision rule $(\gamma_1, \gamma_2) = (0.010{,}0.036)$ presented in Fig. 10.2, $\alpha(H_0^{2/2})$ is controlled at the level of 0.036 while $\alpha(H_0^{1/2})$ is controlled at the level of 0.000625.

4. It is important to emphasize that $\alpha(H_0^{K/K})$ has different meaning from $\alpha(H_0^{1/K})$. Take multiple studies as an example, where the error rate of wrongly approving

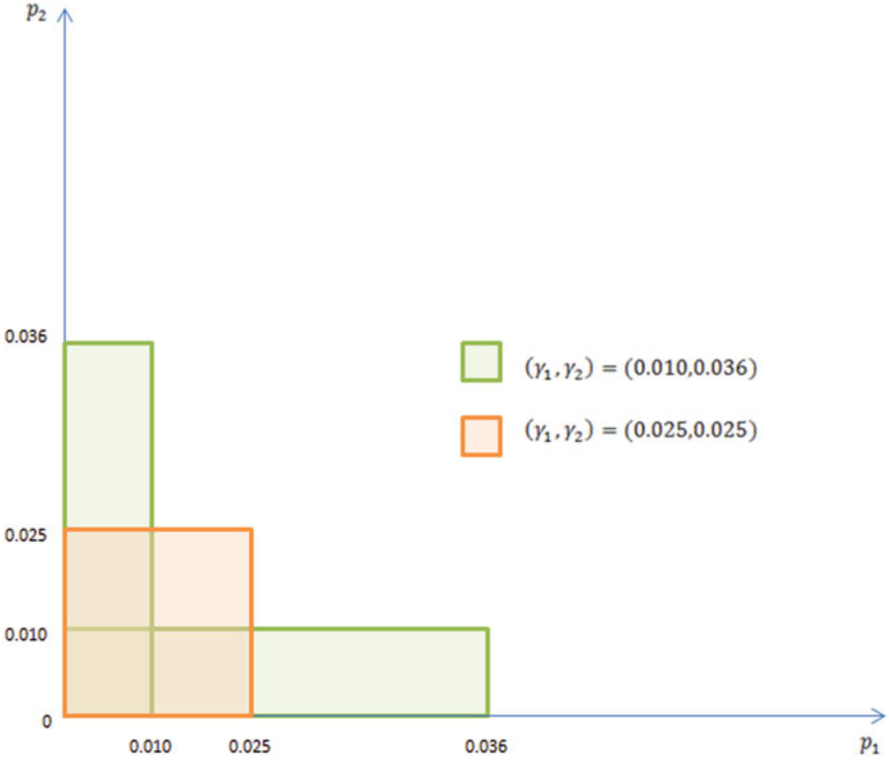

Fig. 10.2 Rejection regions that control the error rate at the level of 0.000625 for two independent studies

> an ineffective drug is controlled at the level of $\alpha(H_0^{1/K})$. The error rate of accepting the hypothesis that all studies present consistent evidence when it is false is controlled at the level of $\alpha(H_0^{K/K})$. It makes common sense that $\alpha(H_0^{1/K})$ should be more stringent in comparison to $(H_0^{K/K})$, as the error of approving an ineffective drug is more serious than the error of claiming consistent evidence when in fact that efficacy is not consistently presented among studies. For two studies, $\alpha(H_0^{1/2})$ is controlled at the level of *0.000625*, while $\alpha(H_0^{2/2})$ is controlled at the level of *0.025* for a decision rule *(0.025, 0.025)*. For two co-primary endpoints, $\alpha(H_0^{1/2})$ is controlled at the level of *0.025*, while $\alpha(H_0^{2/2})$ can be controlled at the level of *0.030* for a decision rule *(0.023, 0.030)*.

The calculation of the decision rules has been described in detail in papers by Li and Huque (Li and Huque [10]) and Li (Li [9]) and various sets of decision rules can be calculated. The original approach of collective evidence requires that the decision rule be prespecified. Since prespecifying a decision rule can be arbitrary and can cause trouble, a practical approach is proposed to reduce the burden of the prespecifying decision rule.

10.3.3 The Practical Approaches to Evaluate Evidence Collectively

To reduce the burden of selecting and prespecifying decision rules, the practical approach uses only one set of decision rules for each K. The following set of decision rules can be considered for independent studies: $(\gamma_1, \gamma_2) = (0.025,\ 0.025)$, $(\gamma_1, \gamma_2,\ \gamma_3) = (0.025,\ 0.025,\ 0.100)$, and $(\gamma_1, \gamma_2,\ \gamma_3,\ \gamma_4) = (0.025,\ 0.025,\ 0.100,\ 0.150)$ for $K = 2,\ 3,$ and $4,$ respectively. The ideal evidence for $K = 2$ is to have two studies demonstrate statistical significance at the same level of 0.025, therefore, the decision rule $(\gamma_1, \gamma_2) = (0.025,\ 0.025)$ should be considered for $K = 2$. The decision rule $(\gamma_1, \gamma_2,\ \gamma_3) = (0.025,\ 0.025,\ 0.100)$ for $K = 3$ is developed from $K = 2$ by adding $\gamma_3 = 0.100$. The choice of γ_3 is primarily driven by controlling $\alpha(H_0^{3/3})$, the error rate of wrongly rejecting the overall null $H_0^{3/3}$, at the level of 0.100. The decision rule for $K = 4$ is similarly derived. Note that, the larger the K is, it is reasonable to accept higher levels of error rates of wrongly rejecting the null $H_0^{K/K}$.

In cases of co-primary endpoints, doses, or tests within one study, the decision rules that are recommended are formed with $\gamma_k = 0.025,\ k = 1,\ 2, \ldots,\ K$. This choice will conservatively control the error rate $\alpha(H_0^{1/K})$, wrongly rejecting $H_0^{1/K}$ in one study, at the level of 0.025. This level of error rate can only be reached when correlation among the co-primary endpoints, doses, or tests is 1. A more realistic level of error rate can be calculated when the range of the correlation can be estimated. The error rate $\alpha(H_0^{K/K})$ is also controlled at the level of 0.025 in one study for the recommended decision rule.

If the p values of the study results satisfy the decision rules, all error rates are adequately controlled. However, it may not be reasonable to require all study results to satisfy the decision rules for drug approval. For example, if the p values of two studies are *(0.020, 0.028)*, this may be considered as convincing evidence for an effective drug. It is therefore necessary to establish the standard of convincing evidence. To address this, two quantities are proposed, one to measure the worst inflation and the other for consistency.

The worst inflation is the maximum possible error that could be observed and is defined in (10.1) below. It is the probability of observing the kth p value $p_{(k)}$ that equals to $\max(\gamma_k, pv_{(k)})$ or less, where $pv_{(k)},\ k = 1,\ 2,\ \ldots,\ K$, are the ordered observed p values. The relative inflation is calculated using formula (10.2).

$$Max.\ Inflated\ error = P\left(\bigcap\nolimits_{k=1}^{K} p_{(k)} \le \max\left(\gamma_k, pv_{(k)}\right)\right) \quad (10.1)$$

$$\%\ of\ Inflation = \frac{P\left(\bigcap_{k=1}^{K} p_{(k)} \le \max(\gamma_k, pv_{(k)})\right) - P\left(\bigcap_{k=1}^{K} p_{(k)} \le \gamma_k\right)}{P\left(\bigcap_{k=1}^{K} p_{(k)} \le \gamma_k\right)} \quad (10.2)$$

For example, if the observed p values of two independent studies are *(0.020, 0.028)*, the inflation is

Table 10.2 Examples of max. inflation and consistency using decision rule (0.025, 0.025)

Observed p values	Max. inflated error (%)	Consistency
0.020, 0.026	0.000675 (8.0 %)	17.0 %
0.021, 0.026	0.000675 (8.0 %)	14.1 %
0.026, 0.026	0.000676 (8.2 %)	0.0 %
0.027, 0.027	0.000729 (16.6 %)	0.0 %
0.025, 0.028	0.000775 (24.0 %)	8.4 %
0.020, 0.030	0.000875 (40.0 %)	28.3 %
0.030, 0.030	0.000900 (44.0 %)	0.0 %
0.010, 0.036	0.001175 (88.0 %)	73.5 %
0.025, 0.036	0.001175 (88.0 %)	31.1 %
0.030, 0.036	0.001160 (101.6 %)	17.0 %

$$Max.\ Inflated\ error = P\left(p_{(1)} \leq 0.025 \cap p_{(2)} \leq 0.028\right) = 0.000775$$

$$\%\ of\ Inflation =$$

$$\frac{P\left(p_{(1)} \leq 0.025 \cap p_{(2)} \leq 0.028\right) - P\left(p_{(1)} \leq 0.025 \cap p_{(2)} \leq 0.025\right)}{P\left(p_{(1)} \leq 0.025 \cap p_{(2)} \leq 0.025\right)} = 24\%.$$

The consistency is another measure that helps assess the variation of the observed results against the decision rule $(\gamma_1, \gamma_2, \ldots, \gamma_K)$. There can be several ways of assessing the consistency. The measure introduced here is the sample variance of the relative ratio of the ordered observed p value $pv_{(k)}$ versus the corresponding component of decision rule γ_k , for $k = 1, 2, \ldots, K$. The ratios are considered as the normalized observed p values by the components of the decision rule. The calculation can be written as (10.3):

$$Consistency = \sqrt{\frac{1}{\mathrm{K}-1}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left(\frac{\mathrm{pv}_{(\mathrm{k})}}{\gamma_{\mathrm{k}}} - \frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\frac{\mathrm{pv}_{(\mathrm{k})}}{\gamma_{\mathrm{k}}}\right)^2} \tag{10.3}$$

For the same example above, the consistency is calculated as:

$$Consistency = \sqrt{\left(\frac{0.028}{0.025} - 0.96\right)^2 + \left(\frac{0.020}{0.025} - 0.96\right)^2} = 22.6\%.$$

Table 10.2 lists the calculation of the inflation and consistency of some observed p values for the case of two independent studies using decision rule (0.025, 0.025).

The collective evidence approach uses one set of predetermined criteria which control the desired level of error rates. To make the decision flexible and evidence based, the evidence obtained from the study calculated as the maximum inflation and consistency are also taken into consideration. An inflation of 24 % with less than 22.6 % consistency for two studies (p values are 0.020, 0.028) may be considered as convincing evidence. However, the approval decision should be determined in conjunction with the disease indication, drug safety profiles, and availability of other drugs for the same indication in the market. Other factors such as the selection of outcome measures and the similarity of study design among the studies included for evaluation may also be factored in for the decision making.

10.4 Collective Evidence in Drug Evaluation

As discussed earlier, it can be helpful to divide the error into different types, i.e., the error of wrongly approving an ineffective drug, the error of wrongly choosing the optimal doses, and the error of false labeling drug information. The first logical step in drug evaluation is to evaluate if a drug is efficacious by controlling the error rate of wrongly approving an ineffective drug. Once it is concluded that the drug is efficacious and reasonably safe, the next step is to identify the optimal doses. Selection of optimal drug doses is not discussed, as it involves evaluating the risk–benefit ratio and possibly pharmacokinetic information which is beyond the scope of this chapter. The discussion of the multiple doses is focused on the efficacy evaluation here. The next step is the labeling decision by controlling the error rate of labeling false information. The error rates are discussed in this section for cases of multiple studies, doses, endpoints, and tests.

10.4.1 Multiple Studies

The total evidence from multiple studies can be obtained by conducting a meta-analyses or using the collective evidence approach. For either approach, the first step is to decide which studies are to be included in the evaluation, since diversely designed studies may not always be informative when evaluating evidence collectively. The studies should be selected based on the study population, design, and the conduct of the studies, rather than the results of studies. It is also important to select studies based on a well-defined patient population. Study design factors, such as blinding, treatment duration, endpoints, and usage of concomitant medications, are important considerations as well. The conduct of the studies, such as the time period and condition of implementation, can be crucial too. For example, studies of seasonal allergic rhinitis may need to be conducted during allergy seasons when high levels of pollen are apparent in the air. If heavy rain occurs, the participants may not be exposed to sufficient allergen to develop allergic reactions. Inadequate exposure could be a legitimate reason to exclude the study, whereas, certain design differences may not be a valid reason to exclude studies. For instance, study endpoints may be evaluated differently among studies in allergic conjunctivitis studies. The redness of the eyes can be evaluated either by study subjects themselves or by practitioners. This may not be a valid reason to exclude studies even though it can be argued that the self evaluation may inherit larger variability than that from the practitioners. To obtain an unambiguous analysis, a good practice is to develop an integrative analysis plan to prespecify criteria for study selection.

The following step is to select a statistical method to evaluate the evidence collectively. Patient-level meta-analyses to poll studies have been popular approaches and are desirable if the number of studies is large and all studies are similarly designed. The collective evidence approach can be desirable in situations when it is important to understand the individual study results and consistency among them; also when differences in study design prohibit study pooling and cause difficulties in interpretation for meta-analyses.

The collective evidence approach for independent studies is relatively easy to use and interpret. To illustrate, take another hypothetical example of a set of p values from three independent studies. The evaluation could be simple if the results satisfy the decision rule for $K = 3$, $(\gamma_1, \gamma_2, \gamma_3) = (0.025, 0.025, 0.100)$. The error rate should be controlled at the level of 0.000156 based on $P(\bigcap_{k=1}^{3} p_{(k)} \leq \gamma_k) = 6\gamma_1\gamma_2\gamma_3 - 3\gamma_3\gamma_1^2 - 3\gamma_2^2\gamma_1 + \gamma_1^3$. If the observed p values were *0.001, 0.020,* and *0.120,* it would be necessary to calculate the inflation and consistency using Formulas (10.2) and (10.3). The % inflation and consistency is 24.0 and 58.9 %, respectively. Suppose that the results were obtained from three studies used to support allergic conjunctivitis and redness was the primary endpoint. Further, assume that the endpoint was assessed by the patients in the study yielding the p value of 0.120 and the other two were assessed by physicians. If patients were less trained, the reporting variability could be larger than the clinician reported outcomes. Hence, the level of inflation and consistency could be considered reasonable for recommending approval. Even if the endpoints were assessed consistently in all three studies, such results might reflect a possible situation that the drug worked for certain patients that were included in the studies, perhaps not consistently. Depending upon the consistency level and the observed maximum p value, it might be useful to further investigate who were more likely to benefit from the drug and who were not.

10.4.2 Multiple Doses to Support Efficacy Evaluation

The evaluation of multiple doses may serve two different purposes: the efficacy evaluation and the selection of the optimal doses. Discussion in this chapter is focused on the efficacy evaluation only.

A typical multiple dose study design includes parallel arms of several doses and placebo where MAPs are traditionally applied. As a result of the stringent significance levels by controlling the type I error, the sample sizes for each arm need to be increased. The application of MAPs could limit the enthusiasm and feasibility to include multiple doses, which are imperative for better understanding of the efficacy as well as dose–response relationship.

It could be suggested that the MAPs for multiple doses in individual studies do not appear to be useful in either efficacy evaluation or the identification of the optimal doses. The fact that all doses show the trend of efficacy is strong evidence against an

Table 10.3 Illustration of strategies of evaluating two doses in two studies using the practical collective evidence approach

	p values		Strategy 1		Strategy 2	
Study	High dose	Low dose	Trend test	Inflate error for studies	Inflated error for dose	Inflated error for studies
1	0.028	0.015	0.023	0.000625 (0 %)	0.028	0.000775 (24 %)
2	0.024	0.013	0.021		0.025	

ineffective drug. It can be even stronger evidence if a reasonable dose–response relationship is demonstrated consistently in multiple studies. Some believe that MAPs are necessary for identifying the effective doses. The counter argument could be made that if a drug is efficacious, many of the dose levels should be efficacious. Whether the dose levels can reach statistical significance is a matter of sample size and treatment difference. Instead of identifying the efficacious doses by the significance, a helpful strategy is to determine the optimal doses, which should be based on the risk–benefit profiles, effect sizes, and other information. If *p* values play any roles in the identification of the optimal doses, the rank of the *p* values is usually sufficient. It is unnecessary to use any adjusted *p* values because the rank of either adjusted or unadjusted *p* values is the same.

An exception to keep in mind is that certain drugs may have a narrow therapeutic window where many doses may not support the efficacy. In those cases, the understanding of the dose–response relationship is more important than adjusting *p* values. The efficacy can then be established by a consistent dose–response relationship in multiple studies.

Evaluation of multiple doses should depend upon the study design. For a typical phase III study, two or three doses that are likely to be the optimal doses are selected based on information from early phase studies. It is expected that all doses would demonstrate efficacy to a certain degree. Two strategies are discussed to evaluate multiple doses collectively. The first one is to use directional tests to establish the efficacy by modeling the dose–response trend. This requires a good understanding of the true dose–response which could be obtained from early phases of clinical studies. Guidance on the directional tests, also referred to as the trend tests, is discussed by Li and Lagakos (Li and Lagakos [11]). When multiple studies are available, the trend tests should be first performed within individual studies. Then the *p* values obtained from the trend tests should be evaluated using the practical collective evidence approach. The second strategy is to evaluate the multiple doses within the individual studies first by using the practical collective evidence approach, then to evaluate the evidence across studies. The two strategies are illustrated in the following hypothetical example, using two studies with two doses in each study. The *p* values of high and low doses of the two studies as well as the results of the two strategies are listed in Table 10.3.

- To illustrate strategy 1, assume from early phase studies that decreased trend was observed as the dose increased. The pseudo-dose indicators were coded as

2 for the low dose and 1 for the high dose. The p values of the trend tests were hypothetical values, 0.023 and 0.021 for studies 1 and 2, respectively. The results satisfied the two-study decision rule $(\gamma_1, \gamma_2) = (0.025{,}0.025)$.

- To illustrate strategy 2, the worst inflated error for multiple doses for each study was calculated first. The decision rule used for two doses was $(\gamma_1, \gamma_2) =$ (0.025,0.025). The worst inflated error due to two doses for study 1 was 0.028, a result by assuming the worst possible correlation between two doses (a very conservative approach). Similarly, the worst inflation for study 2 was 0.025. The worst inflation for two studies was then calculated as 0.000775 with 24 % inflation. The consistency was 8.4 %. If the correlation between two doses was known or can be estimated, the inflated error could be calculated relatively accurately and should be smaller than that presented in Table 10.3. It can be concluded that the evidence of efficacy is convincing.

If MAPs are applied to this example, depending upon the choice of the procedure, it is likely that results in study 1 are considered inconclusive. The statistical decision rules across studies are unavailable.

Strategy 1 should also be considered when many doses are included in a single study, such as phase II dose ranging studies. Strategy 1 should be particularly useful for drugs with narrow therapeutic windows where a nonlinear dose–response trend could be specified.

10.4.3 Multiple Endpoints

Diseases are multifaceted entities where one endpoint is usually insufficient to describe a certain aspect of a disease or reflect disease changes. Therefore, multiple endpoints are used in clinical studies. Endpoints are chosen based on the study objectives, usually the indications for drugs. For example, a drug approved for chronic obstruction pulmonary disease (COPD) can have indications as a bronchodilator, to reduce exacerbation, or to prolong survival. Each indication is evaluated by a set of prespecified endpoints. The endpoints are usually organized as the primary, secondary, and exploratory endpoints in the study protocols. The primary endpoints are defined by the medical communities, including the FDA, and are crucial for the approval of drug indications. The selection of the secondary endpoints is relatively flexible and may depend upon the secondary objectives or features relevant to the primary endpoints and a particular drug. The exploratory endpoints may be less relevant to the indication and often are included in the study for other purposes.

In efficacy evaluations, the primary endpoints must demonstrate clinically and statistically significant benefits in order for the indication to gain approval. The secondary endpoints should be supportive of the primary endpoints by showing trend of treatment benefit. Clearly, the primary and secondary endpoints play different roles and have different expectations in efficacy evaluation. The natural hierarchical order among the different types of endpoints implies that the secondary endpoints would not contribute any additional error in the efficacy evaluation for a specific indication. Hence, no MAPs are needed because of the hierarchical structure of the different types of endpoints for evaluating a specific indication.

In the labeling process, the primary endpoints of the approved indication are always described in the label. What is less clear is the selection of the secondary endpoints. Again, because of the natural hierarchical order, MAPs are not needed for the different types of endpoints in labeling process. However, MAPs may be considered for the multiple secondary endpoints in labeling. This is discussed in detail in a later section.

10.4.4 Co-Primary Endpoints

Often more than one primary endpoint is used to evaluate a disease condition. European Medicines Agency (EMEA) (Committee for Proprietary Medical Products (CPMP) by EMEA [4]) requires all co-primary endpoints to be statistically significant at the level of one-sided 0.025. The limitation of this approach is that as more co-primary endpoints are used, it becomes more difficult to show all endpoints statistically significant. The ordinary least squares (OLS) and generalized least squares (GLS) tests proposed by O'Brien (O'Brien [15]) consider consolidating all co-primary endpoints into one test. Another practice is to develop one composite primary endpoint by combining all co-primary endpoints. The problems of the composite endpoints are widely discussed in the literature (Kip et al. [8]; Montori et al. [13]). The main problem of the O'Brien's OLS and GLS tests as well as the composite endpoints is that they may disguise the heterogeneity in treatment responses among the co-primary endpoints.

The approach of collective evidence is similar to EMEA's approach which emphasizes the understanding of individual performance of all co-primary endpoints. The collective evidence approach simply recognizes the room of flexibility when controlling the error rate of wrongly rejecting the null hypotheses. When there are multiple studies, similar to the case of multiple doses, the collective evidence of the co-primary endpoints is to first calculate the maximum inflated error within each study. Using the maximum inflated error of the individual studies, the maximum inflated error for all studies is then calculated as well as the consistency index. Again, a hypothetical example is used to illustrate the application in a scenario of two studies using two co-primary endpoints. The p values as well as the results of applying the practical collective evidence approach are listed in Table 10.4.

10.4.5 Secondary Endpoints

This section focuses on the discussion of controlling the error rate of labeling false information due to multiple secondary endpoints. Often, the statistically significant

Table 10.4 Illustration of evaluating two co-primary endpoints using the practical collective evidence approach

Study	Primary 1	Primary 2	Inflated error of co-primary endpoints at the level of 0.025	Inflated error of two studies at the level of 0.000625
1	0.023	0.030	0.030	0.000875 (40 %)
2	0.018	0.025	0.025	

secondary endpoints are labeled. With such practice, the more endpoints that are evaluated, the higher chance to show statistical significance. For this reason, it may be necessary to use MAPs to control the error rate of labeling false information, however, not within individual studies when multiple studies are available.

Without loss of generality, the case of two studies is illustrated. Suppose that both studies evaluate Endpoints A and B. Let A_1 and A_2 represent the results of Endpoint A from study 1 and study 2, respectively, and B_1 and B_2 for Endpoint B from study 1 and study 2, respectively. If a MAP is used in the individual studies, the logic should be written as

$$(A_1 \cup B_1) \cap (A_2 \cup B_2) = A_1A_2 \cup A_1B_2 \cup A_2B_1 \cup B_1B_2$$

The logic controls the error rate for four possible outcomes A_1A_2, A_1B_2, A_2B_1, and B_1B_2 that have the potential to become statistically significant or positive, when they are in fact false. With a close look of the four possible outcomes, it is only possible to claim A_1A_2 or B_1B_2, as they represent the situations where the same endpoint is significant in both studies. The outcomes A_1B_2 and A_2B_1 would never be considered in the label in reality as they represent the cases that endpoint A is significant in one study as well as B is significant in the other. Thus, it is unnecessary to control the error that would never be committed.

Alternatively, if each endpoint is first evaluated across studies collectively, the only possible outcomes are A_1A_2 or B_1B_2. Then it makes sense to apply MAPs to control error due to the two possible outcomes to make claim. For instance, if the error rate of labeling false information should be controlled at the level of 0.025 and there are ten secondary endpoints, applying the Bonferrion correction, each endpoint should be controlled at $\alpha(H_0^{1/K}) = 0.0025$ for K studies. Notice that it is not recommended that the level of error rate for the secondary endpoints be as stringent as the error rate of wrongly approving an ineffective drug. The mistake of wrongly approving an ineffective drug is a more serious matter than that of labeling a false endpoint. In practice, p values that are significant at level of 0.025 consistently across studies are labeled, which is more stringent than necessary. Hence, the adjustment with MAPs may not be necessary unless the number of secondary endpoints is in the scale of hundreds and more.

Often, it is useful to order the secondary endpoints based on the clinical importance in the integrative statistical analysis plan. This is equivalent to using the closed testing procedure on the secondary endpoints. So, the clinically more relevant endpoints are labeled if there is consistently convincing evidence across studies.

It can be further debated whether only placing the significant secondary endpoints in drug labels is an efficient way of communicating drug information. For clinically important secondary endpoints, the statistically insignificant results may be as important to share as the significant ones with patients and practitioners. Insignificant results may inform practitioners that the drug has not shown convincing evidence on certain clinically important secondary endpoints.

10.4.6 Multiple Tests

In this chapter, multiple tests are referred to as performing multiple analyses on the same endpoint and using the same set of data, which is different from the multiple tests for different endpoints, such as gene analyses. Multiple tests are commonly used in clinical studies and usually structured as the primary analysis and secondary analyses (or sensitivity analyses). The primary endpoints are often analyzed using multiple methods, usually with the prespecified primary analysis in an intent-to-treat (ITT) population and several secondary analyses. The multiple tests are used to ensure a good understanding of the treatment benefit from the primary analysis and relatively consistent evidence across all tests.

It is important that all the primary and secondary analyses should be valid and reasonable analyses. Valid analyses are unbiased under null. Reasonable analyses are those that the power under alternative is not seriously distorted and the treatment benefit is not overly underestimated or exaggerated. For example, baseline-carry-forward is sometimes used in missing data imputation and a valid analysis under null. However, this approach may not be a reasonable analysis as it could be overly conservative and the test result would be biased towards null if the treatment is to prevent disease from deterioration. In other scenarios, the approach could exaggerate the treatment difference if the disease symptoms can be improved over time without treatment. The worst-case-carry-forward approach is another valid test under null; however, it is not considered reasonable, as it could overly exaggerate the treatment differences under alternatives in certain scenarios. Another valid test is the test for proportions. It may not be a reasonable test when there are differential dropouts between treatment arms, perhaps due to toxicities.

It may not be equitable that the primary analysis is the most powerful analysis or the only important analysis in making conclusion. This is particularly true when handling missing outcome data. Often missing outcome data are missing-not-at-random and there is not one imputation approach that is better than others. The good practice is to prespecify one imputation method for the primary analysis. Multiple methods, served as sensitivity analyses, are used to confirm that the result of the primary analysis does not deviate from other imputation methods too much and that the impact of the missing data is small. In addition, the totality of evidence obtained from multiple tests may enhance the understanding of treatment difference. For example, the family of weighted log-rank tests and the proportional hazard model are all similarly structured (Harrington and Fleming [6]) and are valid tests under null, but can be sensitive to different types of treatment differences revealed in the data. The commonly used log-rank test, the unweighted test, is more sensitive to differences manifested later than the Wilcoxin log-rank test which is more sensitive to differences exhibited earlier. The discussion here is not to undermine the importance of prespecifying the primary analysis. Prespecifying one primary analysis is particularly important when reporting the results in publications and drug labels. The rule of thumb is to report the primary analysis, rather than by picking the best results among all analyses, while all analysis results should be taken into consideration for decision making and interpretation.

It can also be argued that multiple tests may not necessarily inflate the type I error rate, given that all tests are reasonable and valid. A valid test has a 0.025 chance to reach statistical significance under null. The chance for the majority of the tests to show statistical significance together cannot be larger than 0.025 under null. Following the principle of collective evidence, it would not be convincing evidence if only one test shows a significant result, while other analyses lack statistical significance. Conversely, the evidence would be considered convincing if the majority of the tests reveal statistically significant (or close to) results.

10.5 Case Studies

10.5.1 Case 1: The Primary Endpoint Failed

All relevant information discussed in this case can be found in the FDA advisory briefing package (US [21]). Spiriva Handihaler (tiotropium) was first approved for maintenance treatment of COPD based on forced exploratory volume in 1 second (FEV_1). In 2009, the sponsor submitted the results of a study titled understanding the potential long-term impacts on function with tiotropium (UPLIFT) seeking several usage indications, among them, COPD exacerbation. UPLIFT was a randomized, double-blinded, and placebo-controlled multicenter study. A total of 5993 COPD patients were randomly assigned to tiotropium or placebo in a 1:1 ratio, 2987 to tiotropium and 3006 to placebo. The patients were treated over a 4-year period. Another 6-month study that was conducted in approximately 2000 COPD veterans (VA) was also available. The exacerbation results of the two studies are summarized in Table 10.5. As can be seen from Table 10.5, the primary endpoint for exacerbation, the time from randomization to the first exacerbation episode, was statistically significant in both studies. The average risk reduction over time in both studies was about 15 % in tiotropium in comparison to placebo. All the secondary endpoints listed in Table 10.5 were statistically significant at the two-sided level of 0.050. Despite the statistically significant results shown in two studies, the approval of the exacerbation indication was debated among FDA's statistical reviewers and in the advisory committee meeting.

Table 10.5 Summary of exacerbation results in the UPLIFT and VA studies

	UPLIFT			VA study		
	Tio $N = 2986$	Placebo $N = 3006$	Ratio (p val)	Tio $N = 914$	Placebo $N = 915$	Ratio (p val)
Median time (month)	16.7	12.5	0.86 (< 0.001)	–	–	0.83 (0.034)
Total # of events	6691	7183	–	376	446	–
Rate (#/p-y)	0.73	0.85	0.86 (< 0.001)	0.71	0.88	0.81 (0.037)
# of exacerbation days/p-y	12.1	13.6	0.89 (0.001)	10.0	12.6	0.79 (0.056)

The complication was that the primary endpoint of the UPLIFT study was the rates of decline in FEV_1. UPLIFT failed to show any difference in rates of decline in FEV_1. Exacerbation was a secondary endpoint in the UPLIFT study. Furthermore, the study prespecified a closed testing procedure requiring that the primary endpoints show statistically significant treatment differences before testing the secondary endpoints.

Following the prespecified decision rule, it was argued that because the primary endpoint failed, the secondary endpoints should no longer be tested for the reason of protecting type I error. Consequently, there was no sufficient evidence for the exacerbation indication.

An opposing view stated that overly emphasizing the prespecified statistical decision rules could be problematic, and the fact that multiple studies were available could have reduced the need to use the decision rule. The prespecified decision rule was not necessarily scientifically valid as it was based on the expectation to the study, which was a hypothesis to be tested. The gambling nature of the prespecified decision rule made the selection appear to be arbitrary. In UPLIFT, the study allowed patients to take any COPD treatments available in the market. The expectation of tiotropium slowing down the deterioration of pulmonary function at the design stage may no longer be valid over the course of the study as COPD treatments evolved over time. Furthermore, when multiple studies were available, the error rate of wrongly approving an indication could be tightly protected.

The advisory committee voted to approve the exacerbation indication. This case exemplified the arbitrary nature of the prespecified decision rules. If the Bonferroni procedure was prespecified, no one would question the efficacy on exacerbation for the exact same study results. The lesson learned is that the evidence-based drug evaluation should not rely on the prespecified decision rule, particularly when multiple studies are available. The collective evidence approach can be useful in post hoc evaluation. In this case, when applying the practical approach proposed in this chapter, as both the UPLIFT and VA studies were statistically significant at the 2-sided level of 0.050, there was no error inflation with the decision rule $(\gamma_1, \gamma_2) = (0.025, 0.025)$.

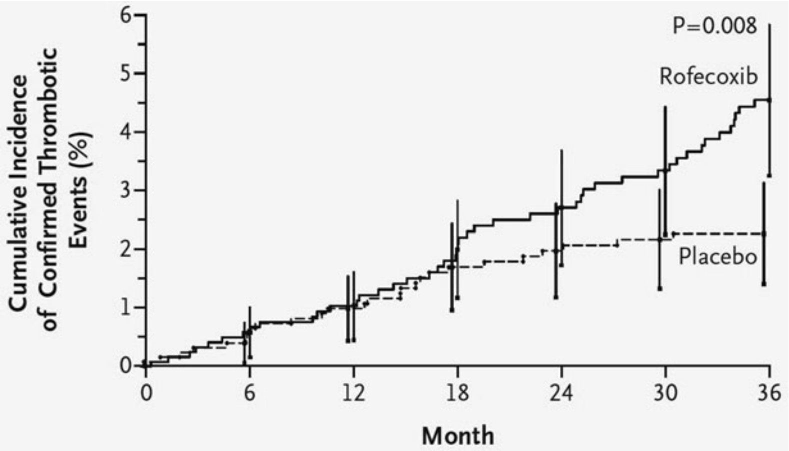

Fig. 10.3 Cumulative CV events curves observed in APPROVe

The error rate of wrongly approving an ineffective drug was indeed tightly protected at the level of 0.000625. The fact that all secondary endpoints were highly significant in both studies further supported the efficacy of tiotropium for treating exacerbation.

10.5.2 Case 2: Dispute on Vioxx (rofecoxib) Cardiovascular Risk

Rofecoxib is a COX-2 agent that was first approved by the FDA in 1999 and withdrawn from market in 2004 due to cardiovascular risk findings in the adenomatous polyp prevention on Vioxx (APPROVe) study (Bresalier et al. [1]). The APPROVe study was a randomized, double-blinded, parallel-grouped, and placebo-controlled study to evaluate the occurrence of neoplastic polyps in patients with a history of colorectal adenomas. Eligible patients were randomized to rofecoxib 25 mg daily or placebo in a 1:1 ratio; 1287 receiving rofecoxib 25 mg and 1299 receiving placebo. At a planed interim analysis, 46 patients developed at least 1 confirmed thrombotic event over 3059 patient-year in the rofecoxib group, and 26 events over 3327 patient-year in the placebo group. The hazard ratio was 1.92 (p value = 0.008) and the cardiovascular risk (CV) risk in rofecoxib was statistically significantly greater compared with placebo. The cumulative incidence curves of the confirmed thrombotic events of the two groups were shown in Fig. 10.3 (Bresalier et al. [1]).

The APPROVe results were published in 2005 in the *New England Journal of Medicine* (NEJM) (Bresalier et al. [1]). In the paper, it was stated that a test of the proportional-hazard (PH) assumption, evaluating the interaction between the treatment and a time logarithm, was specified in the statistical analysis plan for analyzing the cardiovascular risk. Based on this test, the p value of the interaction was statistically significant (two-sided p value = 0.010). It was therefore concluded that the CV risk between the two groups was not proportional over time. Additional post hoc analyses indicated that the CV risk was evident after 18 months of rofecoxib treatment, whereas the CV risk was similar between rofecoxib and placebo for the first 18 months of treatment.

Later, the investigation team reported to NEJM that an error had been identified when reporting the test for the PH assumption (Business Wire [2]) in the original publication (Bresalier et al. [1]). The reported result used linear time rather than the time logarithm that was specified in the analysis plan. The test using the time logarithm yielded a 2-sided *p* value of 0.07, which failed to reach statistical significance at the 2-sided level of 0.05. However, Merck insisted that using linear time was an appropriate analysis based on their diagnostic tests. Therefore, their conclusion of CV risk after 18 months would be unchanged (Business Wire [2]).

NEJM issued a correction (NEJM [14]) in 2006 indicating that the prespecified test using a time logarithm should be the correct analysis. As this analysis did not reach statistical significance, the PH assumption was not rejected. Therefore, a conclusion about the CV risk of rofecoxib should not be made for treatment after 18 months.

An important lesson learned from this case is the interpretation of multiple tests of the PH assumption. Both tests, using linear time or logarithm of time, are valid and reasonable tests. The prespecified test is not necessarily the best test. On the other hand, it is a good statistical practice to report data using the prespecified test. In disputing the PH assumption, although the test using a time logarithm does not reach statistical significance at the two-sided level of 0.05, a *p* value of 0.07 was considered marginally significant. Adding the evidence from the test using linear time, which was statistically significant, the totality of evidence demonstrated that the CV risk ratio was not constant over time. However, the fact that risk ratio was not constant over time did not infer the absence of the CV risk in the first 18 months of the rofecoxib treatment. The interaction tests simply could not answer if rofecoxib caused harm in the first 18 months of treatment.

It is important to reemphasize that the collective evidence approach is not to abandon the prespecification and planning. On the contrast, careful planning and designing experiments, prespecifying the experiment procedures, hierarchy of endpoints, the primary analyses, and all other secondary or sensitivity analyses, as well as safety measures and evaluation are imperative for achieving scientific rigor. However, throughout the discussion of the chapter, prespecifying a decision rule in a study appears to add more confusion in drug evaluation.

10.6 Remarks

Drug evaluation is a complex process that involves multidisciplines including medical, drug safety, statistical, clinical pharmacology, chemistry, and preclinical reviews. The decision is based on collective evidence from all disciplines, a different level of synthesizing evidence collectively. Still, drug efficacy is the key element, as none of the other evaluations would be necessary if a drug was ineffective. This explains why there have been significant efforts to develop statistical methodologies to define systematic approaches to control the error of wrongly approving an ineffective drug. The collective evidence approach is an effort to enrich and improve the systematic approaches.

The collective evidence approach reintroduces the "AND" logic which has been overlooked in drug evaluation. With this foundation, the approach takes all available evidence in decision making, controls various errors occurring in drug evaluation, balances the need for consistency among evidence, and allows reasonable variation. The proposed practical approach may reduce the burden of arbitrarily selecting prespecified decision rules in the individual study protocols. It is noteworthy that this approach does not relax the standard of drug approval; rather it provides an alternative way of evaluating evidence with proven scientific rigor.

Rigidly using the collective evidence approach can also be problematic. As discussed earlier, drugs having narrow therapeutic windows may not have multiple doses supporting the efficacy. However, the collective evidence approach can be applied to examine if a consistent dose–response relationship is exhibited in multiple studies. The application of the collective evidence approach may need special care in drug safety evaluation as well. The safety evaluation usually takes a less conservative approach. On one hand, the risk signal that occurred in one study or one dose can be valuable information for practitioners and patients. On the other hand, a trend of risks consistently occurring in multiple studies, albeit statistically insignificant, can raise serious concerns.

This discussion does not cover the multiplicity issues occurring in subgroup analyses, multiregion studies, and interim analyses. The problems noted in such situations may not all be simple multiplicity problems. Nevertheless, the principles of the collective evidence approach can be applied in evaluating evidence when these multiplicity issues occur.

For future research in the area of collective evidence, utility function can be an alternative approach to summarizing evidence collectively. Eriksen and Keller (Eriksen and Keller [5]) proposed a quantitative way of combining evidence from clinical efficacy and safety data to preclinical safety data of drugs using utility function. This idea can be extended to combine multiple endpoints, multiple doses, and multiple studies. More research needs to be done to further develop this approach.

Acknowledgment The author would like to thank a referee's thoughtful comments and constructive suggestions.

References

Bresalier RS, Sandler RS, Quan H (2005) Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. NEJM 352(11):1092–102

Business Wire (2006) Merck corrects description of a statistical method used in APPROVe study—study results unchanged. http://www.businesswire.com/news/home/20060-

530005860/en/Merck-Corrects-Description-Statistical-Method-APPROVe-Study. Accessed 30 May 2006

Chuang-Stein C, Stryszak P, Dmitrienko A et al. (2007) Challenge of multiple co-primary endpoints: a new approach. Stat Med 26:1181–1192

Committee for Proprietary Medical Products (CPMP) by EMEA (2001) Points to consider on multiplicity issues in clinical trials. Biometrical J 43:1039–1048

Eriksen S, Keller R (1993) A multi-attribute-utility-function approach to aeighing the risks and benefits of pharmaceutical agent. Med Decis Mak 13:188–125

Harrington DP, Fleming TR (1982) A class of rank test procedures for censored survival data. Biometrika 69:133–143

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802

Kip KE, Hollabaugh K, Marroquin OC et al. (2008) The problem with composite end points in cardiovascular studies. J Am Coll Cardiol 51(7):701–707

Li QH (2009) Evaluating co-primary endpoints collectively in clinical trials. Biometrical J 51(1):137–145

Li QH, Huque MF (2003) A decision rule for evaluating several independent clinical trials collectively. J Biopharm Stat 13:621–628

Li QH, Lagakos SW (2006) On the Relationship between directional and omnibus statistical tests. Scand J Stat 33:239–246

Marcus R, Peritz E, Gabriel KR (1976) On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63:655–660

Montori VM, Permanyer-Miralda G, Ferreira-González I et al. (2005) Validity of composite end points in clinical trials. BMJ 330:594–596

NEJM (2006) Correction on cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. NEJM 355:221

O'Brien PC (1984) Procedures for comparing samples with multiple endpoints. Biometrics 40:1079–1087

Pocock SJ (1997) Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. Control Clin Trials 18(6):530–545

Proschan MA, Waclawiw MA (2000) Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials 2:527–539

Shih WJ, Quan H (1999) Planning and analysis of repeated measures at key time-points in clinical trials sponsored by pharmaceutical companies. Stat Med 18:961–973

US FDA (1995) Statement regarding the demonstration of effectiveness of human drug products and devices. Fed Regist 60:39180–39181 (Docket No. 9500230) Accessed 1 Aug 1995

US FDA (1998) Guidance for industry providing clinical evidence of effectiveness for human drug and biological products. http://www.fda.gov/downloads/Drugs/.../Guidances/ucm078749.pdf. Accessed May 1998

US FDA (2009) Pulmonary-allergy drugs advisory committee package. http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/drugs/pulmonary-allergydrugs-advisorycommittee/ucm190463.pdf Briefing package. Accessed 19 Nov 2009