Clin Drug Investig (2015) 35:133–140
DOI 10.1007/s40261-014-0263-5

ORIGINAL RESEARCH ARTICLE

Interpreting the Results of a Retrospective Comparison of Test and Reference Treatments in a Randomized Clinical Trial Setting

Moshe Fridman · M. Haim Erder

Published online: 20 December 2014

Abstract

Background and Objectives The retrospective comparison of test and reference treatment arms in a randomized prospective clinical trial is potentially useful in economic modeling seeking to assess the cost effectiveness of alternative therapies.

Methods To enhance the credibility of such retrospective comparisons, we propose the application of the following adjustments to significance levels obtained from standard statistical methodology: (1) a significance test for the lower bound of the 95 % confidence interval for the observed difference, (2) a conservative Bonferroni method of adjustment for multiple comparisons, (3) an adjusted *p*-value calculated using Scheffe's single-step method, and (4) Bayesian 95 % credibility intervals with a prior centered at zero.

Results These adjustments were applied to data from a randomized double-blind concurrent trial (SPD489-325) that established the efficacy and safety of lisdexamfetamine dimesylate (LDX) in children and adolescents with attention-deficit/hyperactivity disorder (ADHD). Prospectively planned analyses demonstrated that the reduction in the symptoms of ADHD was significantly greater than placebo in patients treated with either LDX or the reference treatment, osmotic-release oral system methylphenidate (OROS-MPH). Retrospective analyses showed that the improvement in the symptoms of ADHD was greater in patients treated with LDX than OROS-MPH. We now show that this observation remained significant after the application of the four statistical penalties.

Conclusions By adjusting the significance level, it is possible to compare quantitatively such retrospective results with prospectively defined comparisons. However, the qualitative level of such retrospective evidence should remain secondary to that obtained from prospectively specified comparisons in a randomized clinical trial.

Key Points

To improve credibility of retrospective comparisons in randomized clinical trials, we propose four statistical methods to discount the observed *p*-value or 95 % confidence interval to account for the retrospective nature of the analysis

Potentially useful retrospective results are currently not available because of a lack of appropriate standardization methodology that can allow comparison to prospective results. The proposed methods provide a tool for such comparisons

1 Introduction

Prospectively posing the research hypotheses along with the design of the experiment, as well as defining the methods for data analysis, are at the heart of scientific research methods. A well-designed clinical trial protocol should clearly state the statistical hypotheses and statistical tests planned, and provide a power analysis to determine the adequacy of the proposed sample size to achieve the

M. Fridman (✉)
AMF Consulting, Inc., Los Angeles, CA, USA
e-mail: fmoshe@amf-consulting.com

M. H. Erder
Shire Development LLC, Wayne, PA, USA

predetermined study goals. The purpose of this prescribed approach is to ensure integrity and prevent bias in the scientific development process. Clinical trials, however, often generate additional information that fall outside the bounds of the planned analyses but that may be of value in clinical, formulary, and reimbursement decision making.

In considering control groups for pivotal clinical trials, regulators typically require the test drug to demonstrate superiority to placebo. In addition, the inclusion of a third arm consisting of a gold-standard reference therapy is considered optimal [1–3]. The purpose of the parallel comparison of a reference treatment with placebo is to provide evidence of the validity and sensitivity of the study design and execution [1–3]. However, once the predefined comparison of the experimental drug with placebo has been conducted in such a three-arm trial, and the efficacy of the reference treatment compared with placebo has established the validity and sensitivity of the study design and execution, the question arises if it is acceptable to compare the test and reference arms retrospectively. A review of clinicaltrials.gov identified 79 interventional, placebo-controlled, phase III studies in any therapy area, completed (with results) since the beginning of 2010 and that included an active reference arm. Thus, clinical trials that include a reference arm as well as test and placebo arms are not uncommon and the retrospective comparison of test and reference therapies is potentially a rich source of clinically useful information. However, when not predefined, such a comparison of active treatment arms is controversial. One concern is the risk of publication bias (i.e., the tendency to publish only positive results), and one might argue that such unplanned retrospective findings should never be accepted for publication. A second concern is that pharmaceutical companies, which are obligated to report study findings per protocol to government agencies, could avoid pre-specifying more 'risky' comparisons to avoid having to report unfavorable findings. However, when the primary endpoint data for the reference treatment are collected under exactly the same conditions as the test drug, a retrospective comparison of active treatment arms may be informative and of value. Indeed, given the cost and delay involved in conducting an additional, head-to-head, randomized clinical trial, the retrospective comparison of the two active treatments in a three-arm clinical trial may provide the only means to compare the test drug with a standard therapy of known efficacy and safety and, when used appropriately, may advance the understanding of how a medicine performs and support decision making by payers seeking to establish policy for the reimbursement of alternative treatments.

In the present paper, we suggest that confidence in retrospective comparisons of active treatments in multi-arm clinical trials may be increased by imposing statistical penalties designed to raise the threshold for such unplanned analyses to be considered 'statistically significant'. We describe four 'penalty' methods and apply them to the retrospective comparison of the two active treatment arms in study SPD489-325 [4, 5]. This was a randomized, double-blind, dose-optimized, placebo-controlled, phase III trial of the prodrug stimulant lisdexamfetamine dimesylate (LDX) in children and adolescents with attention-deficit/hyperactivity disorder (ADHD), which included a reference arm of the standard therapy, osmotic-release oral system methylphenidate (OROS-MPH).

2 Methods

2.1 Adjusting Significance Levels from Standard Retrospective Comparisons

It is assumed that the active treatment arms to be compared retrospectively were part of a randomized and well-controlled clinical trial, and that the reference treatment was a standard therapy of known efficacy and safety. Intuitively, when analyzing clinical trial results retrospectively, the probability of a type I error should be anywhere between the achieved p-value, ignoring the fact that the test was not prospectively defined, and the maximal value of one. Stating this from a confidence level viewpoint, the level of confidence we require from this result should be at least 95 % (as it would be if the test was proposed prospectively). This logic suggests that, in addition to clearly labeling results that were retrospectively proposed, the strategy for interpreting this evidence should be based on a penalty for its retrospective nature. We describe four methods of 'adjusting' the significance level and confidence in the observed difference between two active treatment arms from randomized clinical trials to account for the retrospective nature of the analyses. Obviously, for the adjustments to be meaningful, the outcome of the conventional analysis must be statistically significant because the adjustments are designed to reduce the level of significance from that observed prior to the adjustments.

2.1.1 Method 1: Significance Test for the Lower Bound of the 95 % Confidence Interval for the Observed Difference

The observed difference and confidence level used (95 %) in the retrospective test defines a confidence interval (CI). Without loss of generality, we assume that the observed point estimate for the difference between treatments is negative (i.e., a negative difference indicates improvement for the test treatment compared with the reference). Then, the upper bound for a negative difference that is closer to zero can be thought of as the 'worst-case scenario' for the observed difference that is consistent with the alternative hypothesis. Assuming that the upper bound of such a 95 % CI is the observed point estimate for the difference, then a one-sided test of significance to assess if the lower bound is less than zero, using the original test statistic denominator will result in a more conservative test than the one that would be prospectively defined.

2.1.2 *Method 2: Simultaneous Testing Procedures*

If the comparative test had been proposed prospectively, it could have been tested at the standard significance level used in the study for the rest of the primary objective tests. However, given that the comparison was proposed retrospectively, this no longer applies and a more stringent significance level is required. One way to recalculate the significance level for this test could be to view this situation as having added a secondary ad hoc test whose results should be adjusted for multiple comparisons using a family-wise level. This method has a built-in objective mechanism to determine how low the new significance level should be to accept the evidence from the retrospective comparative test.

There are several options for a single-step family-wise adjustment of the significance level. We suggest the highly conservative Bonferroni method [6]. This method requires that the family-wise error is divided amongst the planned comparisons, e.g., a three-arm study with three possible pairwise comparisons tested at 5 % (two-sided) would require an adjustment such that each pairwise comparison is compared with $p = 0.017$.

2.1.3 *Method 3: Adjusted p-Values*

Results from the simultaneous test procedures suggested in Method 2 may be more intuitively understood if, instead of adjusting the significance level, adjusted p-values are reported [7]. This method is similar to Method 2 and works by restating the unadjusted p-value based on its relationship to the adjusted significance level. For example, for the single-step Bonferroni (B) and Sidak (S) methods we have

$$P(B)_{\mathrm{adj}} = n \times P_{\mathrm{unadj}} \qquad (1)$$

and

$$P(S)_{\mathrm{adj}} = 1 - \left(1 - P_{\mathrm{unadj}}\right)^{n}, \qquad (2)$$

where n is the number of hypotheses being tested. For Scheffe's single-step method, the unadjusted p-value is found by calculating the ratio of the comparison sum of squares ($\mathrm{SS_c}$) over the mean square error and finding the tail of an F distribution with 1 and $(N-g)$ degrees of freedom, where g is the number of tests and N the total sample size. The adjusted p-value is found by calculating the F statistic as the ratio of $\mathrm{SS_c}/(g-1)$ over the mean square error and finding the tail of an F distribution with $(g-1)$ and $(N-1)$ degrees of freedom. Multistage procedures may be applied to obtain more powerful results than the classic procedures. For this study, the authors have chosen Scheffe's single-step method.

2.1.4 *Method 4: Bayesian 95 % Credibility Intervals*

The Bayesian method proposed by Matthews can be used to assess quantitative credibility, taking explicit account of prior insights and experience. In our case, the prior information can be used as a penalty for the retrospective nature of the comparison. To this end, we shall apply a prior distribution consistent with the null hypothesis of 'no difference between treatments'. For convenience, a normal prior will be used for the primary comparison between the test and reference drugs. Mean and variance parameters for the prior normal distribution may be obtained from published results on comparisons between the drugs. As there is no prior information in the literature on the difference between the study and reference arm drugs in our example, we illustrate the method using a prior normal distribution centered at zero (no difference between the drugs) and also find the threshold prior mean value needed to maintain a significant advantage for the study drug. The same variance for the difference between the study and reference arms found in study SPD489-325 is used in the absence of a better prior estimate.

Given an observed 95 % CI for the treatment difference

$$[L_{\mathrm{D}},\ U_{\mathrm{D}}] = \bar{X}_{\mathrm{D}} \pm 1.96\ S_{\mathrm{D}} \qquad (3)$$

Bayes's Theorem provides the means of combining evidence captured as a prior distribution. Using a normal prior distribution, the result is a posterior distribution expressed in the form of a credible interval $[L_{\mathrm{P}}, U_{\mathrm{P}}]$ in which

$$[L_{\mathrm{P}},\ U_{\mathrm{P}}] = \bar{X}_{\mathrm{P}} \pm 1.96\ S_{\mathrm{P}} \qquad (4)$$

with S_{P} and $\bar{X}_{P}$ calculated from

$$(1/S_{\mathrm{P}}) = (1/S_{0}) + (1/S_{\mathrm{D}}) \qquad (5)$$

and

$$(\bar{X}_{\mathrm{P}}/S_{\mathrm{P}}) = (\bar{X}_{0}/S_{0}) + (\bar{X}_{\mathrm{D}}/S_{\mathrm{D}}) \qquad (6)$$

For the results to be credible, the posterior 95 % credibility interval $[L_{\mathrm{P}}, U_{\mathrm{P}}]$ should still show an advantage for the test drug.

2.2 SPD489-325: A Randomized, Double-blind, Placebo- and Active-controlled Clinical Trial

To explore the implications of the above penalty methods, we applied them to results obtained from a randomized, double-blind, placebo- and active-controlled clinical trial (SPD489-325) of LDX in children and adolescents with ADHD [4, 5]. The study was conducted in accordance with current applicable regulations and the standards of good clinical practice. The primary endpoint was the change from baseline to endpoint in ADHD symptoms measured using the ADHD Rating Scale IV (ADHD-RS-IV) total score [8]. This scale is derived from the 18 inattentive and hyperactive/impulsive diagnostic criteria for ADHD in the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition [9]. The range of the scale is 0–54 and a reduction in score indicates an improvement in ADHD symptoms. The study was powered to show a difference between each active treatment and placebo and the pre-specified comparisons were between LDX and placebo and between OROS-MPH and placebo, adjusted for baseline ADHD-RS-IV total score, age group (6–12, 13–17 years), and country (nine European countries). Least-squares means were estimated for each of the treatment arms and for the differences between treatment arms using an analysis of covariance model. A formal statistical test was not pre-specified between LDX and OROS-MPH. The randomization and blinding procedures, data collection and monitoring, data double-entry, logical checks, and querying were all done prospectively, in a uniform manner irrespective of the fact that the comparison between the two active drugs was not planned in the protocol, and with the same level of scrutiny in accordance with the study sponsor's standard operating procedures. OROS-MPH, the reference treatment in this study, is a long-acting methylphenidate formulation of well-established efficacy and safety [10, 11], and is approved for the treatment of children, adolescents, and adults with ADHD [12].

3 Results

In the primary outcome from study SPD489-325, reductions in the ADHD-RS-IV total score were significantly greater in patients treated with LDX ($N = 104$) or OROS-MPH ($N = 107$) than in those who received placebo ($N = 106$) [4]. The estimated least-squares mean change (standard error) in ADHD-RS-IV scores from baseline to study endpoint for the LDX, OROS-MPH, and placebo groups were −24.30 (1.16), −18.72 (1.14), and −5.70 (1.13), respectively [4]. The planned comparisons between each treatment arm and placebo were statistically significant and are provided in Table 1. When tested retrospectively, it was found that the improvement in symptoms of ADHD was significantly greater for LDX than OROS-MPH (Table 1) [5]. The impact of the four methods for adjusting the outcome of this retrospective comparison will now be assessed.

3.1 Method 1: 95 % CI Lower Bound

The 95 % CI bound for the LDX vs OROS-MPH difference in adjusted mean change from baseline ADHD-RS-IV score was −8.45, −2.70. Using the upper bound of the 95 % CI to represent the worst-case scenario for the mean difference between active treatment groups, the new *p*-value from a significance test that the upper bound mean difference (−2.70) is less than zero is 0.034. This result supports a statistically significant improvement in patients receiving LDX compared with OROS-MPH.

3.2 Method 2: Simultaneous Testing Procedure

Using the conservative Bonferroni method of adjustment for multiple comparisons, we obtain a *p*-value significance cut-off of 0.017 instead of 0.05. With an observed *p*-value of 0.0002 for the retrospective comparison of the two active treatment arms, a statistically significant improvement in patients receiving LDX compared with OROS-MPH is still supported in the presence of the adjustment.

3.3 Method 3: Adjusted *p*-Values

The adjusted *p*-value calculated using Scheffe's single-step method was found by calculating the F statistic as the ratio of $SS_c/2$ over the mean square error and finding the tail of an F distribution with (2) and (302) degrees of freedom. In this case, the adjusted *p*-value equals 0.027, indicating that the improvement in patients treated with LDX was statistically significantly greater than in those who received OROS-MPH.

3.4 Method 4: Bayesian 95 % Credibility Intervals

In SPD489-325, the 95 % CI for the difference between LDX and OROS-MPH in least-squares mean ADHD-RS total score was −8.45, −2.70 [5]. Using a normal prior distribution with a mean of zero and a standard deviation of 1.46, we calculated the posterior standard deviation S_P to be 0.73 and the posterior 95 % credibility interval to be −4.22, −1.35, indicating that LDX was significantly more effective than OROS-MPH despite this adjustment. The threshold value for the prior distribution mean difference between LDX and OROS-MPH was 2.75, namely an ADHD-RS-IV total score advantage of OROS-MPH by 2.75 units.

Table 1 Change in ADHD-RS-IV total score from baseline to study endpoint[a]: comparison between treatment arms in study SPD489-325 [4, 5]

Comparison	Difference in least-squares means	Standard error	95 % CI for difference	*p*-value
LDX vs. placebo (planned)	−18.60	1.456	(−21.47, −15.74)	<0.0001
OROS-MPH vs. placebo (planned)	−13.03	1.436	(−15.85, −10.20)	<0.0001
LDX vs. OROS-MPH (retrospective)	−5.57	1.460	(−8.45, −2.70)	0.0002

ADHD-RS-IV ADHD Rating Scale IV, *CI* confidence interval, *LDX* lisdexamfetamine dimesylate, *OROS-MPH* osmotic-release oral system methylphenidate

[a] Endpoint was defined as the last on-treatment post-baseline study visit with a valid ADHD-RS-IV total score

3.5 Comparison of Methods

In general, Method 1 would result in the strictest penalty for conducting the comparison of the test and reference treatments retrospectively. Based on this method, we first create a 95 % CI for the mean difference between treatments and consider the CI bound closer to zero, that is a shift from the CI midpoint of about 1.96 standard errors. The penalty involves shifting another 1.64 standard errors towards zero by testing the difference of this new point from zero using a one-sided test. This is equivalent to requiring a minimum difference of 3.6 standard errors from zero for the observed mean difference between treatments, or a significance level of about 0.0002.

Methods 2 and 3 will generally coincide in their conclusion, as Method 2 calculates a penalized significance level and Method 3 calculated an adjusted *p*-value based on similar multiple statistical comparisons methodology. For the typical three-arm clinical trial, the resulting cut-off *p*-value is 0.017, not nearly as restrictive as Method 1.

For Method 4, the degree of penalty can vary depending on the prior distribution used. Clearly, the more distant towards the opposite side of zero the prior mean is relative to the observed mean difference between treatments, the smaller the prior variance value and the larger the sample size for the source of the prior information, the stronger the prior evidence against the observed results and the stricter the penalty. Figure 1 presents a sensitivity analysis of the credibility intervals as the mean and standard deviation of the assumed prior distribution are varied.

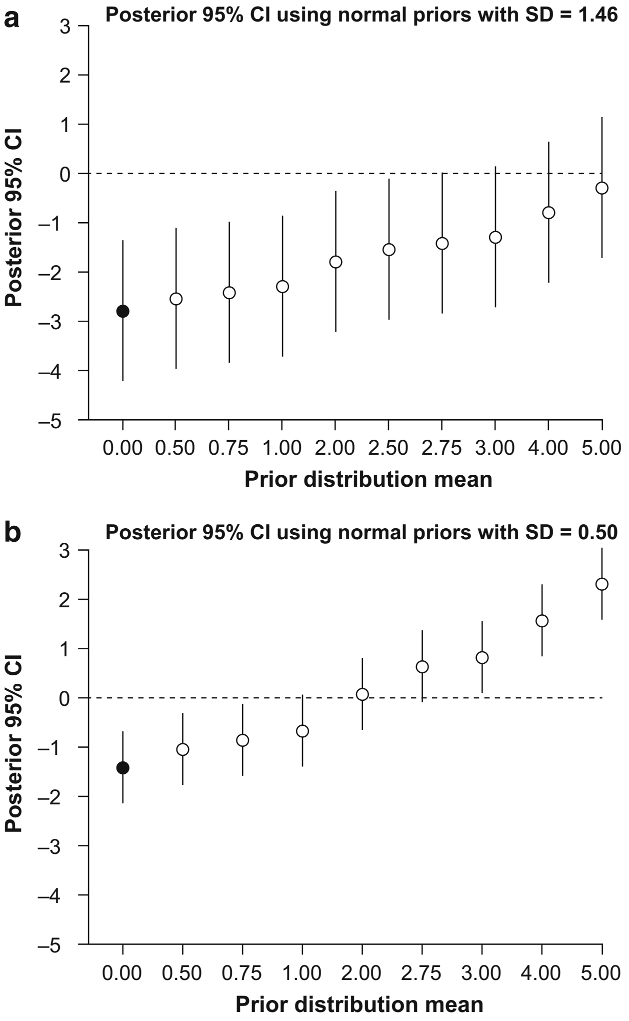

Fig. 1 Sensitivity analysis of the 95 % Bayesian credibility intervals (Method 4) over mean and standard deviation of the prior distribution. *CI* credibility interval, *SD* standard deviation

4 Discussion

We propose that the application of statistical penalties to the outcomes of a retrospective comparison of test and reference therapies in a three-arm clinical trial will add to the credibility of the analysis when used to aid decision making by, for example, formularies and payers. The unplanned retrospective comparisons of the two active treatments in a randomized, double-blind, dose-optimized, placebo- and active-controlled, phase III trial of LDX in children and adolescents with ADHD (SPD489-325) concluded that the reduction in symptoms was statistically significantly greater in the LDX group than in the active OROS-MPH group [5]. We now report that the difference between the two active treatments remains statistically significant when four different statistical penalties are applied, suggesting that the findings from this retrospective comparison are robust.

Although controversial, retrospective analyses of experimental data can yield useful information beyond the predefined objectives of a study. One approach to the retrospective analysis of clinical trial data is a meta-analysis, a standard well-established method for the integration of summary statistics (e.g., effect size, standard errors, and sample sizes) across trials. Typically, such analyses pool results from similarly designed, comparative clinical trials. However, one of the major problems in combining findings remains the weighting of data to reflect the accurate 'value of information' from each study. To address this issue, DerSimonian and Laird [13] examined eight published meta-analyses and proposed a method to assign trial weights using a random effect size approach. A second approach is the post hoc analysis of clinical trial data. Such analyses may be used to discover new indications, particularly for unsuccessful compounds [14], combine patient data from several clinical trials or apply other stratification schemes [15], and combine clinical trial patient data with other sources such as historical data [16]. In the above examples, clinical trial information is used to address questions that are external or broader than the goals of the original trials and are, therefore, accepted as legitimate reuse of the data.

In the specific case of study SPD489-325, the comparison of test and reference arm was not pre-specified. However, the clinical trial was conducted in a manner that would have allowed for the comparison of the LDX and OROS-MPH treatment arms had that comparison been pre-specified. The reference arm data used in the retrospective comparison are of equal quality to those of the protocol-specified analyses because the study was a randomized double-blind trial such that treatment was unknown. No modifications were made to the design or conduct of this study because this test was not planned, and all of the required information has been collected and is available. Moreover, the retrospective comparison of test and reference therapies is informative because it includes all of the information about the treatment effect available in the study sample. For these reasons, we argue that the methods of analysis that should have been used to compare active treatments had the comparison been prospectively planned can still be applied. This is in marked contrast to the retrospective statistical analyses of subgroups in which the post hoc definition of subgroups of interest can be a source of bias, outcomes are strongly influenced by the size of the subgroups, and which usually involve multiple hypothesis testing [17–19].

To improve the credibility of the retrospective statistical comparison of treatment arms in clinical trials such as SPD489-325, we suggest that the results of the analysis be discounted compared with results obtained from hypotheses tested prospectively. To achieve this, we have proposed four statistical methods to penalize the observed p value or 95 % CI to account for the retrospective nature of the analysis. In the example of study SPD489-325, the retrospective comparison of test and reference therapies indicated that symptomatic improvements based on ADHD-RS-IV total scores were significantly greater in patients receiving LDX than OROS-MPH [5]. This comparison remained statistically significant when each of the four penalty methods were applied: (1) using a significance test for the upper bound of the 95 % CI for the observed difference ($p = 0.034$), (2) using a conservative Bonferroni method of adjustment for multiple comparisons ($p < 0.017$), (3) using an adjusted p value calculated using Scheffe's single-step method ($p = 0.027$), and (4) using a Bayesian 95 % credibility interval with a prior centred at zero (-4.22, -1.35). These results are strongly supportive of statistical significance for the retrospective comparison of LDX and OROS-MPH in study SPD489-325.

For results to remain significant after adjustment, the unadjusted comparison should typically be highly significant. In other words, the results should be indicative of a strong difference in treatment efficacy or based on a large sample or both. Clearly, the impact of the penalties will depend on how strongly significant was the observed retrospective p value: the four methods suggested here are dependent on the mean and standard error of the observed difference between the test treatment and the reference treatment arms. In addition, Methods 2 and 3 depend on the number of additional multiple comparisons conducted while Method 4 depends on the prior distribution specification. Given the conceptual difference in the approach for applying a penalty on the observed significance of the retrospective comparison between Method 1, Methods 2 and 3, and Method 4, the conclusions from the methods may diverge, particularly when the observed retrospective comparison p-value is only moderately significant. There is no single gold standard method or sequence of methods we recommend using and for a particular situation only a single method of penalized testing needs to be applied. In our example, we applied all four methods for illustrative purposes. The choice of the method should be transparent. The choice between a frequentist or Bayesian analysis method is subjective and may depend on the availability of reliable and useful prior information. Choosing between a more conservative approach such as Method 1 or less restrictive ones such as Methods 2 or 3 may depend on the level of risk involved with the decision, similar to the dilemma regarding the determination of a type I error level.

The difficulties that arise with the retrospective statistical comparison of data obtained from prospective clinical trials are publication bias and the interpretation of the findings. Regarding publication bias, clinical trial sponsors are required to report on the results of the prospectively approved treatment comparisons, but not on unplanned retrospective ones. For the type of trial with a test treatment, reference treatment, and placebo arms, sponsors are required to report on the comparisons with the placebo arm. Because there is no standard framework for publication of the retrospective comparison of test and reference treatments, these results are currently not likely to be published, which may lead the reader to conclude, perhaps erroneously, that there was no difference between test and reference treatment arms. If the suggested penalty methodology is accepted it will provide a platform for the publication of such comparisons.

As for the interpretation of such findings, this type of statistical evidence is a hybrid between a prospective controlled clinical trial and a retrospective analysis, in which the data are from the former and the inference is from the latter. In this scenario, because the data are obtained in a controlled randomized setting, they are as reliable as in standard randomized clinical trials. However, it is clear that we have a lower level of confidence in the significance of retrospective comparisons than those obtained a priori planned comparisons using controlled clinical trial data. Results obtained with unplanned comparisons should always be disclosed as such and judged with caution. We suggest that the qualitative value of the evidence obtained from such hybrid comparisons be classified between evidence from prospective randomized clinical trials and that from retrospective studies with non-randomized data.

Although we argue that unplanned comparisons of active treatment arms from a randomized clinical trial can be tested credibly, especially when statistical outcomes are discounted to account for their retrospective nature, it is important to consider any potential clinical or biological caveats of performing such a comparison. Do, for example, the selected dose(s) and frequency of administration, study duration, or study outcomes unduly favor one treatment over another? In the case of SPD489-325, the authors of the original undiscounted retrospective comparison of the active treatment arms took care to discuss whether the fact that the maximum permitted dose of the reference therapy OROS-MPH in the European countries in which the trial was conducted was lower than that permitted in North America may have impacted the outcome of the comparison [5].

5 Conclusions

Concerns related to the retrospective efficacy analyses include publication bias and the possibility that only positive results will be widely reported. However, we conclude that the retrospective comparison of active treatment arms in a randomized double-blind trial is meaningful when the data used in the retrospective comparison are of equal quality to those of the protocol-specified analyses and when methods designed to penalize the classical confidence or significance level are employed. Of course, the qualitative level of such retrospective evidence should remain secondary to that obtained from prospectively formulated comparisons from a randomized clinical trial.

Acknowledgments The authors gratefully acknowledge Dr. Matthias Bischof from Shire Development, LLC for his helpful methodological suggestions. They also thank Dr. Elizabeth Gandhi and Dr. Eric Southam of Oxford PharmaGenesis^{TM} Ltd for providing editorial assistance, including collating the comments of the authors and editing the manuscript for submission, funded by Shire Development LLC.

Conflict of interest Dr. Fridman is an employee of AMF Consulting, Inc., and served as a consultant to Shire Development, LLC. Dr. Erder is an employee of Shire Development, LLC.

References

1. Food and Drug Administration. Guidance for industry. E10 Choice of Control Group and Related issues in Clinical Trials. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073139.pdf (2001). Accessed 30 Nov 2014.
2. European Medicines Agency. Reflection paper on the need for active control in therapeutic areas where the use of placebo is deemed ethical and one or more established medicines are available. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/01/WC500100710.pdf (2011). Accessed 30 Nov 2014.
3. Koch A, Rohmel J. Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. J Biopharm Stat. 2004;14(2):315–25. doi:10.1081/BIP-120037182.
4. Coghill D, Banaschewski T, Lecendreux M, Soutullo C, Johnson M, Zuddas A, et al. European, randomized, phase 3 study of lisdexamfetamine dimesylate in children and adolescents with attention-deficit/hyperactivity disorder. Eur Neuropsychopharmacol. 2013;23:1208–18. doi:10.1016/j.euroneuro.2012.11.012.
5. Soutullo C, Banaschewski T, Lecendreux M, Johnson M, Zuddas A, Anderson C, et al. A post hoc comparison of the effects of lisdexamfetamine dimesylate and osmotic-release oral system methylphenidate on symptoms of attention-deficit hyperactivity disorder in children and adolescents. CNS Drugs. 2013;27:743–51. doi:10.1007/s40263-013-0086-6.
6. Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1994.
7. Wright SP. Adjusted *P*-values for simultaneous interference. Biometrics. 1992;48:1005–13.

8. DuPaul GJ, Power T, Anastopoulos AD, Reid R. ADHD Rating Scale-IV: checklist, norms and clinical interpretation. New York: Guildford Press; 1998.

9. Diagnostic and Statistical Manual of Mental Disorders. 4th ed. Washington, DC: American Psychiatric Assocation; 1994.

10. Coghill D, Seth S. Osmotic, controlled-release methylphenidate for the treatment of ADHD. Expert Opin Pharmacother. 2006;7(15):2119–38. doi:10.1517/14656566.7.15.2119.

11. Banaschewski T, Coghill D, Santosh P, Zuddas A, Asherson P, Buitelaar J, et al. Long-acting medications for the hyperkinetic disorders: a systematic review and European treatment guideline. Eur Child Adolesc Psychiatry. 2006;15(8):476–95. doi:10.1007/s00787-006-0549-0.

12. CONCERTA Extended-Release Tablets Product Information. http://www.concerta.net/adult/prescribing-information.html (2013). Accessed 30 Nov 2014.

13. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88 pii:0197-2456(86)90046-2.

14. Cavalla D, Singal C. Retrospective clinical analysis for drug rescue: for new indications or stratified patient groups. Drug Discov Today. 2012;17(3–4):104–9. doi:10.1016/j.drudis.2011.09.019.

15. Fleischmann RM, Baumgartner SW, Tindall EA, Weaver AL, Moreland LW, Schiff MH, et al. Response to etanercept (Enbrel) in elderly patients with rheumatoid arthritis: a retrospective analysis of clinical trial results. J Rheumatol. 2003;30(4):691–6 pii:0315162X-30-691.

16. Whitstock MT, Pearce CM, Ridout SC, Eckermann EJ. A retrospective analysis of VIOXX in Australia: using clinical trial data and linked administrative health data to predict patient groups at risk of an adverse drug event. Aust N Z J Public Health. 2010;34(4):431–2. doi:10.1111/j.1753-6405.2009.00579.x.

17. Sormani MP, Bruzzi P. Reporting of subgroup analyses from clinical trials. Lancet Neurol. 2012;11(9):747. doi:10.1016/S1474-4422(12)70181-3 (author reply -8).

18. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine: reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357(21):2189–94. doi:10.1056/NEJMsr077003.

19. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365(9454):176–86. doi:10.1016/S0140-6736(05)17709-5.