Pain 139 (2008) 485–493

PAIN

www.elsevier.com/locate/pain

Review and recommendations

Analyzing multiple endpoints in clinical trials of pain treatments: IMMPACT recommendations

Dennis C. Turk [a,*], Robert H. Dworkin [b], Michael P. McDermott [b], Nicholas Bellamy [c], Laurie B. Burke [d], Julie M. Chandler [e], Charles S. Cleeland [f], Penney Cowan [g], Rozalina Dimitrova [h], John T. Farrar [i], Sharon Hertz [d], Joseph F. Heyse [e], Smriti Iyengar [j], Alejandro R. Jadad [k], Gary W. Jay [l], John A. Jermano [m], Nathaniel P. Katz [n], Donald C. Manning [o], Susan Martin [p], Mitchell B. Max [q], Patrick McGrath [r], Henry J. McQuay [s], Steve Quessy [t], Bob A. Rappaport [d], Dennis A. Revicki [u], Margaret Rothman [v], Joseph W. Stauffer [w], Ola Svensson [x], Richard E. White [y], James Witter [z]

[a] *Department of Anesthesiology, University of Washington, P.O. Box 356540, Seattle, WA 98195, USA*
[b] *Department of Anesthesiology, University of Rochester, Rochester, NY, USA*
[c] *Mayne Medical School, University of Queensland, Brisbane, Queensland, Australia*
[d] *United States Food and Drug Administration, Silver Spring, MD, USA*
[e] *Epidemiology, Merck & Co., Blue Bell, PA, USA*
[f] *Department of Symptom Relief, MD Anderson Cancer Center*
[g] *American Chronic Pain Association, Rocklin, CA, USA*
[h] *Clinical Research, Allergan, Inc., Irvine, CA, USA*
[i] *Center for Clinical Epidemiology & Biostatistics, University of Pennsylvania, Philadelphia, PA, USA*
[j] *Lilly Corporate Center, Eli Lilly & Co, Indianapolis, IN, USA*
[k] *Center for Global Health, University of Toronto, Toronto, Ontario, Canada*
[l] *Schwarz Biosciences, Research Triangle Park, NC, USA*
[m] *NeurogesX, Inc., San Carlos, CA, USA*
[n] *Analgesic Research, Needham, MA, USA*
[o] *Celgene Corporation, Summit, NJ, USA*
[p] *Pfizer, Inc., Ann Arbor, MI, USA*
[q] *Department of Anesthesiology, University of Pittsburgh, Pittsburgh, PA, USA*
[r] *Department of Psychology, Dalhousie University, Canada*
[s] *Pain Relief, Oxford University, Oxford, UK*
[t] *GlaxoSmithKline, Research Triangle Park, NC, USA*
[u] *United Biosource Corporation, Bethesda, MD, USA*
[v] *Johnson & Johnson, Raritan, NJ, USA*
[w] *Alpharma, Piscataway, NJ, USA*
[x] *AstraZeneca R&D, Sodertalje, Sweden*
[y] *Endo Pharmaceuticals Inc, Chadds Ford, PA USA*
[z] *United States Food and Drug Administration, now at United States National Institutes of Health, USA*

Received 12 March 2008; received in revised form 11 June 2008; accepted 30 June 2008

* Corresponding author. Tel.: +1 (206) 616 2626; fax: +1 (206) 543 2958.
E-mail address: turkdc@u.washington.edu (D.C. Turk).

doi:10.1016/j.pain.2008.06.025

Abstract

The increasing complexity of randomized clinical trials and the practice of obtaining a wide variety of measurements from study participants have made the consideration of multiple endpoints a critically important issue in the design, analysis, and interpretation of clinical trials. Failure to consider important outcomes can limit the validity and utility of clinical trials; specifying multiple endpoints for the evaluation of treatment efficacy, however, can increase the rate of false positive conclusions about the efficacy of a treatment. We describe the use of multiple endpoints in the design, analysis, and interpretation of pain clinical trials, and review available strategies and methods for addressing multiplicity. To decrease the probability of a Type I error (i.e., the likelihood of obtaining statistically significant results by chance) in pain clinical trials, the use of gatekeeping procedures and other methods that correct for multiple analyses is recommended when a single primary endpoint does not adequately reflect the overall benefits of treatment. We emphasize the importance of specifying in advance the outcomes and clinical decision rule that will serve as the basis for determining that a treatment is efficacious and the methods that will be used to control the overall Type I error rate.

Keywords: Multiple endpoints; Multiplicity; Clinical trials; Treatment outcomes; Chronic pain; Acute pain; Sampling error; Type I error

1. Introduction

To facilitate meta-analyses and systematic reviews of clinical trials of pain treatments, the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) has recommended a set of core outcome domains [53] and measures [18], strategies for developing improved measures [54], and methods for determining clinical importance of changes in outcome measures [19]. Based on reviews of the literature and consensus discussions, six core outcome domains were recommended for consideration for chronic pain clinical trials: pain, physical functioning, emotional functioning, participant ratings of global improvement, symptoms and adverse events, and participant disposition [53]. The use of multiple outcome measures has also been recommended for evaluations of the efficacy of treatments for other chronic pain conditions, for example, rheumatoid arthritis [20], osteoarthritis [5], low back pain [14], and neuropathic pain [10]. More generally, multiple patient-reported, clinician-rated, laboratory test, and device measurement endpoints are often used in evaluations of treatment impact for diverse clinical conditions, and are commonly reported in product labeling [61].

A major concern with conducting multiple tests of significance (often referred to as the problem of "multiplicity") of different endpoints in a clinical trial involves the so-called Type I error, the probability that a null hypothesis is rejected when the null hypothesis is actually true. The greater the number of statistical tests performed, the greater the probability that one or more of them will yield a statistically significant result by chance alone. One consequence of conducting multiple analyses is that it increases the likelihood of false positive results, making it possible for an investigator to choose the most favorable result from among many analyses that have been performed [9]. For example, if individual statistical tests, each using a significance level (α) of 0.05, are performed for four specific measures recommended by IMMPACT for the pain, physical functioning, emotional functioning, and participant-rated global improvement outcome domains [18], the chance of falsely rejecting at least one null hypothesis of no treatment difference is 18.5%. This example assumes that the four measures are uncorrelated, which is very unlikely for measures of these outcome domains, and the problem of multiplicity is reduced when the outcome measures are positively correlated.

Because multiple endpoints are often necessary to adequately evaluate the benefits of pain treatment [5,10,14,20,53], consideration must be given to controlling the overall probability of a Type I error and the risk of false positive conclusions in designing clinical trials of the efficacy and effectiveness of pain treatments. Regulatory agencies [9,55], biostatisticians [2], CONSORT guidelines [1], and scientific journals often advocate the use of appropriate adjustments to control the overall probability of a Type I error when multiple endpoints are included in clinical trials, and a single primary endpoint does not adequately reflect diverse benefits of treatment. The objective of this article is to discuss multiplicity and describe strategies for minimizing the risk of false positive conclusions in pain clinical trials with multiple efficacy endpoints.

2. Consensus meeting procedure

An IMMPACT consensus meeting was held that included an international group of 33 participants from universities, governmental agencies, a patient self-help organization, and the pharmaceutical industry. Participants were selected on the basis of their research, clinical, or administrative expertise relevant to the design and evaluation of pain treatment outcomes. An attempt was made to include broad representation of various disciplines and expertise while limiting the size of the meeting to promote frank and efficient discussion. To ensure that all attendees were familiar with the recent

advances in addressing multiplicity, six articles reviewing important issues and strategies involving multiple analyses and endpoints were circulated prior to the meeting [7,12,16,23,39,44]. In addition, background lectures were presented at the meeting that examined (1) general issues regarding multiple endpoints and multiple analyses in clinical trials (JFH, TP, Lemuel A. Moyé, III), (2) responder analyses and state attainment criteria in studies of rheumatic diseases (NB), and (3) regulatory perspectives on multiple endpoints (LB).

3. Classification of endpoints

The primary objectives of most clinical trials include evaluating whether a treatment provides clinical benefit in a sample drawn from a population to which the results will be generalized. Clinical benefit should be defined and assessed as unambiguously as possible because it provides the basis for determining whether the results of the clinical trial have demonstrated evidence of treatment efficacy. The procedure for determining whether the results of the trial have demonstrated efficacy, which has been termed the "clinical decision rule" [7], must be specified prior to beginning data analyses.

Adherence to the recommendation that multiple outcome measures should be used in chronic pain clinical trials to adequately evaluate clinical benefit [53] will involve multiple analyses and, as a consequence, the possibility of an increased risk of false positive conclusions for one or more of the outcome measures. This must be addressed in the design of the clinical trial and in its statistical plan, which must specify whether statistically significant improvements for one, several, or all of these endpoints are required for the trial to have demonstrated clinical benefit of the treatment. Although IMMPACT recommended multiple outcome domains and measures [18,53], minimal guidance was provided regarding whether these should be primary, co-primary, or secondary endpoints, and no attention was paid to methods for addressing multiplicity in pain clinical trials (i.e., the clinical decision rule for interpreting results for multiple endpoints). In addition, many chronic pain clinical trials do not clearly specify the clinical decision rule and which endpoints are primary and secondary; without this information it is often impossible to determine whether a treatment has convincingly demonstrated efficacy relative to a control condition [7].

3.1. Primary endpoint

The primary endpoint in a clinical trial has been defined as "the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial" [55]. The primary endpoint will typically determine whether the study results are considered positive, negative, or uninformative concerning the effect of treatment, regardless of the results for other endpoints. Moreover, the sample size, statistical power (i.e., the probability of rejecting the null hypothesis given that the treatment is actually efficacious), and other features of the clinical trial design will be based on the primary endpoint. The primary endpoint is usually a single coherent measure, which can either consist of a single item or a composite of many individual measurements (e.g., from a questionnaire). The rationale for the selection of the primary endpoint should be included in the protocol, and the use of a reliable and valid outcome measure with which experience has been gained in previous research is strongly recommended.

In recommendations of chronic pain outcome domains and measures [18,53], it has been emphasized that the determination of which endpoints are considered primary and which are secondary depends on the specific treatment objectives of the clinical trial. Before beginning the analysis of the data, investigators should be sure to carefully specify the study hypotheses that provide the basis for the selection of the primary endpoint(s) and the clinical and statistical decision rules for data analysis and interpretation. In clinical trials of analgesics, the primary endpoint will almost always be a measure of pain intensity or pain relief, although pain quality and other aspects of pain could also be assessed. In a regulatory context, when there is a single prespecified primary efficacy endpoint and all additional endpoints are declared as providing only supportive or exploratory information – for example, with respect to identifying additional improvements in physical or emotional functioning that may be a consequence of the treatment – adjustment for multiplicity will typically not be necessary [9]. There are also other circumstances in which multiplicity adjustment is usually not considered necessary, for example, when additional endpoints are used only to explore treatment mechanisms, to examine secondary hypotheses [12], or to generate hypotheses for future study.

3.2. Multiple primary endpoints

Different approaches have been used to specify a clinical decision rule for trials that have more than one primary endpoint. Significant results can be required for each of several primary endpoints to consider a trial "positive." When significant results are required for all the primary endpoints, no adjustment for multiplicity is necessary. Requiring each of multiple endpoints to be significant at the same significance level used for a single primary endpoint reduces the statistical power of the trial (this has been termed the "reverse" multiplicity problem), with the reduction in power being greater with larger numbers of primary endpoints and lower correlations among the endpoints [34,42,44].

When a significant result is required for only one of multiple primary endpoints in order to consider a trial positive – for example, either pain intensity or pain relief – each endpoint must be tested with a significance level that has been corrected for multiplicity. Most commonly, this correction is intended to strongly control the familywise or experimentwise Type I error probability, that is, the probability of erroneously rejecting the null hypothesis for at least one endpoint, regardless of which and how many of the individual hypotheses are true. For example, a Bonferroni-corrected significance level of $0.05/K$, where K is the number of primary endpoints, preserves the familywise Type I error probability at 0.05. When significant results are required for more than one but not all of multiple primary endpoints for a trial to be considered positive, correction for multiplicity is also necessary, and this must take into account the total number of endpoints and the number required for the trial to be considered positive [47].

3.3. Secondary endpoints

A variety of types of secondary endpoints have been used in clinical trials. As discussed by D'Agostino [12], these include variables that (1) provide background and greater understanding of the primary endpoint(s); (2) are separate components of a composite primary endpoint; (3) are important given the treatment's objectives but for which the study does not have adequate power; (4) can aid in understanding the mechanisms of action of the treatment; (5) relate to secondary hypotheses that are not major objectives of treatment; and (6) are intended for exploratory analyses.

As these different uses of secondary endpoints demonstrate, such endpoints can provide additional characterization and understanding of treatment effects, but by themselves are not sufficient to confirm that the treatment is efficacious. There has been controversy with respect to whether it is valid to formally analyze secondary endpoints when the primary endpoint has not demonstrated a statistically significant benefit of treatment (unless specified in advance by the trial's clinical decision rule with adjustment for multiplicity). Such analyses of secondary endpoints are typically (but not always) disregarded in regulatory considerations [22,40,44]. However, given the large and potentially valuable amounts of data that are now collected in clinical trials, it has been argued that methods must be developed for appropriately analyzing major secondary endpoints [12], and various methods have been proposed for this purpose, as discussed below in Section 6.

3.4. Exploratory endpoints

Exploratory endpoints are typically not viewed as being directly related to the primary objectives of a clinical trial but are thought to provide potentially worthwhile information about the treatment or clinical condition being studied that could serve to generate hypotheses for future study. For this reason, endpoints that are prespecified in the design of a clinical trial as exploratory do not require any correction for multiplicity. Exploratory endpoints have also been defined as including those endpoints identified on a post-hoc basis or during interim or final analysis phases of the clinical trial [42]; as long as they are clearly identified as such in the clinical study report, these endpoints also do not require any correction for multiplicity.

3.5. Composite endpoints

Composite outcome measures in which multiple endpoints are combined into a single variable have been used to address a variety of issues in clinical trials [23,42]. Composite endpoints are useful when the disease has many manifestations, all of which are important to consider with respect to summarizing patient outcome. They can also be used to increase the statistical power to detect the effects of treatment. One situation in which composite endpoints are commonly used is when there are multiple events of interest, some of which are rare. For example, a composite endpoint in a cardiovascular trial might be the time from randomization to the first occurrence of *either* myocardial infarction, stroke, or death, the rationale being that the sample size requirement for a trial with time to death as the primary endpoint would be prohibitively large due to the rarity of this event. Another circumstance in which composite endpoints have been used is to avoid multiplicity when several endpoints are thought to be essential to adequately characterize the beneficial effects of a treatment. A prominent example of this approach is the use of the ACR-20 in clinical trials of rheumatoid arthritis [21]. This is a "responder index" in which a patient is considered to be a responder if there is a 20% improvement in tender and swollen joint counts and in three of five additional measures (i.e., patient and physician global ratings of improvement, pain, disability, and an acute phase reactant). Another type of composite endpoint is the sum or average of standardized scores across different but relevant outcome domains, as exemplified by the Multiple Sclerosis Functional Composite [11], which combines the results of tests of ambulatory function, arm function, and cognitive function. Ideally, components of a composite outcome should be biologically related but not too highly correlated (otherwise, a single primary endpoint would be more appropriate).

The disadvantages of composite endpoints are as follows: (1) they generally permit only global, not component-specific, conclusions and are subject to misinterpretation [23]; (2) different components may have different degrees of importance; (3) a treatment

effect may be restricted to a single component (or few components) of the composite; and (4) treatment effects may be qualitatively different for different components of the composite [23,45]. Such endpoints can thus mask a beneficial effect, lack of effect, or even harmful effects for one or more of the components of the composite. For this reason, it is generally recommended that analyses of each component of the composite also be reported when results for composite endpoints are presented. When examined this way, the components of the composite can be considered secondary endpoints that are intended to clarify interpretation of the composite, and no adjustment for multiplicity is required. However, if definitive conclusions about the effects of treatment on individual components of the composite are intended, then this must be specified in the protocol, and adjustment for multiple analyses would be necessary.

It is important to note that several of the measures recommended by IMMPACT for chronic pain clinical trials – including the Brief Pain Inventory [BPI,8], Multidimensional Pain Inventory [MPI, 28], Beck Depression Inventory [BDI,4], and Profile of Mood States [POMS, 37] – can be considered composite measures. However, because the reliability and validity of the total and subscale scores of these measures are well established, these measures have been considered single outcome measures in the IMMPACT recommendations [18] and in this manuscript.

4. Approaches for addressing multiplicity in clinical trials with multiple endpoints

A wide variety of approaches have been recommended for addressing multiplicity in clinical trials and for ensuring that the probability of a Type I error is kept within acceptable bounds [e.g., 42,47,56]. Depending on the nature of a study and its objectives – for example, proof-of-concept vs. confirmatory clinical trials, analyses of efficacy vs. analyses of safety, interchangeability vs. hierarchy of endpoints – different approaches will typically be required [47]. Regardless of which approach is used, however, the selected procedure must be specified in the clinical trial protocol and statistical plan before undertaking any analyses of the data.

4.1. Bonferroni and related stepwise procedures

There are a number of *p*-value-based approaches that can be used to adjust for the analyses of multiple endpoints, which include the Bonferroni test and various improvements designed to increase its power [48]. These procedures have been widely used, mainly due to their simplicity and wide applicability, and each has its own advantages and disadvantages. The Bonferroni test, which is the most well known and simplest of the procedures, involves testing the significance of a treatment effect separately for each endpoint and declaring a treatment effect statistically significant for a particular endpoint if the p value is less than α/K, where K is the total number of endpoints (i.e., statistical tests performed).

Related stepwise approaches include the Holm [26], Hochberg [25], and Hommel [27] procedures. The Bonferroni test has the least power of all these procedures, followed by the Holm, Hochberg, and Hommel procedures in that order [7], although strong control of the familywise error rate is not guaranteed for the Hochberg and Hommel procedures [15]. Bootstrapping and other resampling methods [51,59,60] can be used to modify these procedures to take into account the correlations among the endpoints and, hence, improve power.

4.2. Global multivariate testing procedures

Several procedures have been proposed for testing the global null hypothesis of no treatment effects on any of the endpoints for the case where the vector of outcomes has a multivariate normal distribution [56]. Hotelling's T^2 test is perhaps the most well-known test for this problem, but it is sensitive to treatment effects that are not of clinical interest (e.g., treatment effects that are opposite in sign for different endpoints). Several alternative procedures have been proposed, such as O'Brien's ordinary least squares and generalized least square tests [43] and their modifications [29,49], and the approximate likelihood ratio test [50]. Procedures applicable to binary outcomes or that relax the assumption of multivariate normality also exist [30,31,48]. These procedures suffer from many of the same disadvantages of composite endpoints noted above. They are of less value in circumstances where treatment effects are inconsistent across the endpoints [48] and they permit only global conclusions, leading to difficulties in interpretation. However, these tests can be incorporated in a closed testing procedure [36] in order to yield conclusions concerning the individual endpoints [32,52].

4.3. Secondary endpoints and prospective allocation of alpha

In considering the analysis and interpretation of secondary endpoints in clinical trials in which the primary endpoint is negative, Davis [13] proposed that the analysis of a primary endpoint could be conducted with a prespecified significance level α, and that each of K secondary endpoints could be tested using a significance level of $\alpha/(K+1)$ (the total number of endpoints or analyses). Prentice [46] suggested that to preserve the experimentwise Type I error rate at α, the primary endpoint could be tested using a significance level of $\alpha/2$, and the secondary endpoints could each be tested using a significance level of $\alpha/2K$.

Moyé [39,41,42] proposed the "prospective alpha allocation scheme" for preserving Type I error rates at acceptable levels when there are multiple endpoints. In this approach, the overall significance level for the study is allocated among the primary and secondary endpoints. Moyé suggested that this experimentwise α be capped at 0.10, and that the significance level for testing the primary endpoint be set at 0.05 to permit adequate statistical power for the primary hypothesis and to maintain consistency with accepted standards of evidence. The remaining 0.05 of α can then be distributed (equally or unequally) among the secondary endpoints in accordance with their importance or statistical power requirements. Prospectively allocating alpha in this manner preserves the experimentwise Type I error rate and makes it possible to consider a treatment efficacious when the null hypothesis is not rejected for the primary endpoint but is rejected for one or more of the secondary endpoints. However, when formulated in this way, the prospective alpha allocation scheme preserves the experimentwise Type I error rate at a higher rate than is customarily accepted.

4.4. Gatekeeping procedures

Multiple endpoints may be tested according to hierarchical or "gatekeeping" procedures that involve the prospective specification of families of null hypotheses that are tested in a sequential manner [3,16,17,57]. The most straightforward application of these procedures to the multiple endpoint problem is the "serial" gatekeeping approach [3,57] in which testing of families (or "gates") of null hypotheses in a prespecified sequence continues only when *all* hypotheses in the previous family have been rejected; otherwise, the procedure stops, and hypotheses in families that have not yet been tested cannot be rejected. Because of this strict hierarchical nature of the testing, once a gate is passed, the subsequent family of hypotheses can be tested using the same overall significance level as that used in testing the preceding gatekeeper family. For example, the significance of the treatment effect for the primary endpoint could be the first hypothesis tested (using a significance level of 0.05), and if (and only if) the null hypothesis is rejected, the most important prespecified secondary endpoint could then be tested (also with a significance level of 0.05), and if (and only if) this second gatekeeping step is passed, the less important secondary endpoints could then be tested as a family (with an overall significance level of 0.05). In this example, each endpoint or family of endpoints in the sequence serves as the gatekeeper for subsequent tests in the hierarchy, with conclusions about each endpoint depending on acceptance or rejection of the null hypotheses in the previous steps.

There are several different gatekeeping procedures that have been proposed for testing multiple hypotheses that are applicable to clinical trials with multiple endpoints for which a hierarchical testing order can be prespecified. These include the so-called "parallel" gatekeeping procedures developed by Dmitrienko et al. [16] using the closed testing principle [36]. In this approach, testing of families of null hypotheses in a prespecified sequence continues when *at least one* null hypothesis in the previous family has been rejected, rather than all null hypotheses in the previous family. Finally, in many clinical trials it is desirable to test a sequence of hypotheses that are neither completely serial nor parallel. For these situations, a "tree-structured" gatekeeping approach can be used, in which prespecified testing of hypotheses is guided by a decision tree with multiple branches that correspond to individual hypotheses or endpoints; this approach has been illustrated with a clinical trial with multiple primary, secondary, and tertiary endpoints and both superiority and noninferiority objectives [17].

Adjustment for multiplicity is not required when moving from one family of hypotheses to the next in the serial gatekeeping approach (i.e., an uncorrected overall significance level can be used for testing each family of hypotheses). However, adjustment is necessary when multiple null hypotheses are being tested within a family that are not required to be significant for testing to proceed to the next family. Different approaches can be used for this adjustment, including some of those discussed above (e.g., weighted Bonferroni, Simes, and resampling-based tests) [16].

The principal advantage of the serial gatekeeping approach is the lack of the need to adjust for multiplicity at each stage of testing, which results in generally higher power for endpoints that are near the top of the hierarchy. This approach is conceptually appealing in situations where a natural hierarchy of the endpoints can be confidently specified based on their importance. This hierarchy, of course, needs to be specified prior to data analysis. A potential disadvantage is the typically low power for endpoints that are near the bottom of the hierarchy. If the relative importance of the endpoints is clear and correctly specified in the hierarchy, however, this should not be a major concern. Parallel gatekeeping procedures are more flexible, resulting in generally higher power for endpoints that are near the bottom of the hierarchy at the expense of lower power for endpoints that are near the top of the hierarchy.

5. Other types of multiple analyses

Multiple analyses are common in clinical trials, and the problem of multiplicity can arise when examining several treatments (e.g., an investigational medication, a comparator medication, and placebo), various dosages of the same treatment, repeated measures of the same endpoint at different follow-up times, interim analyses,

subgroup analyses, or combinations of these [7,35,42]. Although it is beyond the scope of this article to discuss corrections for multiple analyses in these very different situations, multiplicity must be considered in all these situations, and many of the approaches we have discussed can be applied.

A critical objective of many clinical trials that also involves multiple analyses is the evaluation of safety data. Multiple endpoints are routinely examined in such analyses, including adverse events, development or exacerbation of disease or symptoms, and changes in vital signs and laboratory or imaging findings. Significance levels are usually not adjusted in these analyses because it is generally considered more important to avoid false negative conclusions about safety findings than to avoid false positive conclusions [9]. Nevertheless, there is a potential for inflation of the rate of false positive safety findings when multiplicity is ignored in safety analyses, and correction for multiple analyses should therefore be considered. Approaches for controlling the proportion of errors resulting from falsely rejecting null hypotheses, which has been termed the "false discovery rate" [6], have been recommended for the evaluation of safety data [38].

6. Recommendations and conclusions

The majority of pain clinical trials collect, analyze, and present multiple outcome measures. As we have emphasized throughout this manuscript, the likelihood of obtaining statistically significant results by chance increases with the number of analyses performed. From a clinical perspective, the most important consequence of reporting multiple analyses without correcting for the increased Type I error rate is that it can be falsely concluded that a treatment has significant benefits when the results are actually due to chance rather than to treatment efficacy.

An important problem in clinical trials involves the concern that investigators might choose to present or emphasize only the most favorable results when multiple statistical analyses have been performed. It is therefore imperative that analyses of clinical trial data be determined by the protocol and a prespecified statistical plan. In general, either one or a limited number of endpoints should be designated as the primary efficacy variable(s) that serve as the basis of the clinical decision rule on which the determination of efficacy will be made. All secondary and exploratory endpoints must also be identified in the statistical plan, along with the specific methods that will be used, if any, to correct for multiplicity. In addition, all reports of clinical trial results must include a complete description of the endpoints included in the trial and the clinical decision rule for determining efficacy. The development of improved methods for addressing multiplicity is an active area of statistical research, including procedures we have not considered, for example, Bayesian approaches [24,58].

When multiple endpoints are examined in clinical trials of pain treatments, it is essential to consider whether methods that correct for multiplicity are needed. In those circumstances in which multiplicity must be addressed, one or more of the approaches described above should be used, as appropriate. Although most of the strategies we have discussed can be used to adjust for multiplicity in different types of pain clinical trials, gatekeeping procedures can be generally recommended because of their ease of application and interpretation in many circumstances, their generally widespread acceptability, and their strong control of the familywise Type I error rate. However, regardless of the specific method selected, we recommend that information about the strategy used to control for multiplicity should be included when registering clinical trials, for example, at http://www.clinicaltrials.gov.

In a placebo-controlled clinical trial of a treatment for chronic pain, for example, the primary endpoint tested would likely be changed in pain intensity. IMMPACT has recommended that physical and emotional functioning and participant reports of global improvement should also be included among the six core outcome domains recommended for chronic pain clinical trials [53]. The primary efficacy analysis could therefore be followed by (1) testing for treatment effects on physical functioning using a single measure of physical functioning – either the BPI Interference Scale [8] or the MPI Interference Scale [28], which assess similar outcomes; then (2) testing for treatment effects on emotional functioning using both the BDI [4] and the POMS [37], which are complementary measures; and then (3) testing for group differences in global improvement. Testing for a treatment effect on physical functioning would be performed only if there was a statistically significant treatment effect for the pain intensity primary endpoint; tests for treatment effects on emotional functioning would be performed only if there was a significant treatment effect for physical functioning, and group differences in global improvement would be tested only if there were significant treatment effects for both the emotional functioning measures (because they are complementary) [18].

In the hypothetical example above, the null hypothesis concerning the pain intensity primary gatekeeper could be tested with an unadjusted significance level of 0.05, the null hypothesis concerning the physical functioning measure could be tested with an unadjusted significance level of 0.05 (because this test is only conducted if the benefit of treatment on the primary endpoint is statistically significant), the null hypotheses concerning the two emotional functioning measures with unadjusted significance levels of 0.05 (because these tests are only conducted if the previous analyses are statistically significant and demonstration of a significant result is required for both to proceed to the final

endpoint), and the null hypothesis concerning global improvement with an unadjusted significance level of 0.05. Of course, the specific endpoints, their sequence, and the approach used for adjusting for multiplicity should be based on the research questions of the clinical trial and the expected benefits of the treatment.

We have emphasized the control of the overall probability of a Type I error when there are multiple endpoints in a clinical trial. There is, however, an inherent tension between Type I error and the so-called Type II error, the failure to conclude that an efficacious treatment is actually efficacious. The complement of Type II error is statistical power, and many of the procedures we have discussed for controlling the overall probability of a Type I error will decrease the statistical power of a clinical trial unless the sample size is increased [33,42]. As Sankoh et al. [48] have emphasized, clinical trials must be designed with the understanding that multiplicity is sometimes unavoidable, that adjustment for its effects must be considered, and that when such adjustments are performed, the sample size must provide adequate statistical power to ensure that meaningful conclusions can be drawn. Although statistical methods for addressing multiplicity reduce the risk of false positive conclusions resulting from chance, they do not substitute for the necessity to clearly pre-specify the study hypotheses, outcomes, and clinical decision rule that will serve as the basis for determining whether treatment is efficacious.

Acknowledgements

The views expressed in this article are those of the authors, none of whom have financial conflicts of interest related to the material presented in this manuscript. No official endorsement by the US Department of Veterans Affairs, US Food and Drug Administration, US National Institutes of Health, or the pharmaceutical companies that provided unrestricted grants to the University of Rochester Office of Professional Education should be inferred.

The authors thank Paul J. Lambiase and Mary Gleichauf for their invaluable assistance in the organization of the IMMPACT meeting. Unrestricted grants were provided to the University of Rochester Office of Professional Education to support the consensus meeting on which this article is based by Allergan, Alpharma, AstraZeneca, Celgene, Cephalon, Eli Lilly, Endo, GlaxoSmithKline, Johnson & Johnson, Merck, NeurogesX, Pfizer, and Schwarz Pharma.

References

[1] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.

[2] Bauer P, Chi G, Geller N, Gould AL, Jordan D, Mohanty S, et al. Industry, government, and academic panel discussion on multiple comparisons in a "real" phase three clinical trial. J Biopharm Stat 2003;13:691–701.

[3] Bauer P, Rohmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. Stat Med 1998;17:2133–46.

[4] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.

[5] Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis: consensus development at OMERACT III. J Rheumatol 1997;24:799–802.

[6] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 1995;57:289–300.

[7] Chi GYH. Multiple testing: multiple comparisons and multiple endpoints. Drug Inf J 1998;32:1347S–62S.

[8] Cleeland CS, Ryan KM. Pain assessment: global use of the brief pain inventory. Ann Acad Med 1994;23:129–38.

[9] Committee for Proprietary Medicinal Products (CPMP). Points to consider on multiplicity issues in clinical trials. London: European Agency for the Evaluation of Medicinal Products, 2002, http://www.emea.europa.eu/pdfs/human/ewp/090899en.pdf.

[10] Cruccu G, Anand P, Attal N, Garcia-Larrea L, Haanpaa M, Jorum E, et al. EFNS guidelines for neuropathic pain assessment. Eur J Neurol 2004;11:153–62.

[11] Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999;122:871–82.

[12] D'Agostino RB. Controlling alphas in a clinical trial: the case for secondary endpoints. Stat Med 2000;19:763–6.

[13] Davis CE. Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. Control Clin Trials 1997;18:557–60.

[14] Deyo RA, Battie M, Beurskens AH, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research: a proposal for standardized use. Spine 1998;23:2003–13.

[15] Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W. Analysis of clinical trials using SAS®: a practical guide. Cary, NC: SAS Institute; 2005.

[16] Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Stat Med 2003;22:2387–400.

[17] Dmitrienko A, Wiens BL, Tamhane AC, Wang X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. Stat Med 2007;26:2465–78.

[18] Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.

[19] Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008;9:105–21.

[20] Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.

[21] Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. The American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.

[22] Fisher LD. Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. Control Clin Trials 1999;20:16–39.
[23] Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? Jama 2003;289:2554–9.
[24] Gönen M, Westfall PH, Johnson WO. Bayesian multiple testing for two-sample multivariate endpoints. Biometrics 2003;59:76–82.
[25] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–2.
[26] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
[27] Hommel G. A stepwise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988;75:383–6.
[28] Kerns RD, Turk DC, Rudy TE. The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). Pain 1985;23:345–56.
[29] Läuter J. Exact *t* and *F* tests for analyzing studies with multiple endpoints. Biometrics 1996;52:964–70.
[30] Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. Biometrics 1993;49:975–88.
[31] Legler JM, Lefkopoulou M, Ryan LM. Efficiency and power of tests for multiple binary outcomes. J Am Stat Assoc 1995;90:680–93.
[32] Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 1991;47:511–21.
[33] Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. J Clin Psychiatry 2004;65:1511–4.
[34] Leon AC, Heo M, Teres J, Morikawa T. Statistical power of multiplicity adjustment strategies for correlated binary endpoints. Stat Med 2007;26:1712–23.
[35] Liu A, Tan M, Boyett JM, Xiong X. Testing secondary hypotheses following sequential clinical trials. Biometrics 2000;56:640–4.
[36] Marcus R, Peritz, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976;63:655–60.
[37] McNair DM, Lorr M, Droppleman LF. Profile of mood states. San Diego: Educational and Industrial Testing Service; 1971.
[38] Mehrotra DM, Heyse JF. Use of the false discovery rate for evaluating clinical safety data. Stat Methods Med Res 2004;13:227–38.
[39] Moyé LA. *P*-value interpretation and alpha allocation in clinical trials. Ann Epidemiol 1998;8:351–7.
[40] Moyé LA. End-point interpretation in clinical trials: the case for discipline. Control Clin Trials 1999;20:40–9.
[41] Moyé LA. Alpha calculus in clinical trails: considerations and commentary for the new millennium. Stat Med 2000;19:767–79.
[42] Moyé LA. Multiple analyses in clinical trials. New York: Springer-Verlag; 2003.
[43] O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079–87.
[44] O'Neill RT. Secondary endpoints cannot be validity analyzed if the primary endpoint does not demonstrate clear statistical significance. Control Clin Trials 1997;18:550–6.
[45] Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. Control Clin Trials 1997;18:530–45.
[46] Prentice RL. On the role and analysis of secondary outcomes in clinical trials. Control Clin Trials 1997;18:561–7.
[47] Sankoh AJ, D'Agostino RB, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Stat Med 2003;22:3133–50.
[48] Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med 1997;16:2529–42.
[49] Tang DI, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. Biometrics 1993;49:23–30.
[50] Tang DI, Gnecco C, Geller NL. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. Biometrika 1989;76:577–83.
[51] Troendle JF. A stepwise resampling method of multiple hypothesis testing. J Am Stat Assoc 1995;90:370–8.
[52] Troendle JF, Legler JM. A comparison of one-sided methods to identify significant individual outcomes in a multiple outcome setting: stepwise tests or global tests with closed testing. Stat Med 1998;17:1245–60.
[53] Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106:337–45.
[54] Turk DC, Dworkin RH, Burke LB, Gershon R, Rothman M, Scott J, et al. Developing outcome measures for pain clinical trials: IMMPACT recommendations. Pain 2006;125:208–15.
[55] U.S. Department of Health and Human Services. Guidance for industry: E9 statistical principles for clinical trials. Rockville, MD: Office of Training and Communication, Food and Drug Administration, 1998.
[56] Wassmer G, Reitmeir P, Kieser M, Lehmacher W. Procedures for testing multiple endpoints in clinical trials: an overview. J Stat Planning Inference 1999;82:69–81.
[57] Westfall PH, Krishen A. Optimal weighted, fixed sequence, and gatekeeping multiple testing procedures. J Stat Planning Inference 2001;99:25–40.
[58] Westfall PH, Krishen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. Stat Med 1998;17:2107–19.
[59] Westfall PH, Wolfinger RD. Multiple tests with discrete distributions. Am Stat 1997;51:3–8.
[60] Westfall PH, Young SS. Resampling-based multiple testing. New York: Wiley; 1993.
[61] Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Control Clin Trials 2004;25:535–52.