Why, when and how to adjust for multiplicity in clinical trials: a perspective on regulatory activities*

Dario Gregori[1], Ileana Baldi[2], Anna Chiara Frigo[1], Vincenzo Guardabasso[3]

[1] Laboratory of Epidemiological Methods and Biostatistics,
Dept. of Environmental Medicine and Public Health, University of Padova.
[2] University of Torino and CPO Piemonte
[3] Policlinico Vittorio Emanuele, University of Catania

Corresponding Author:
Prof. Dario Gregori
Laboratory of Epidemiological Methods and Biostatistics
Department of Environmental Medicine and Public Health
Via Loredan 18 - 35131 Padova, Italy
Phone: +39 02 00612711
Fax: +39 02 700445089
Email: dario.gregori@unipd.it

Summary
The need for multiplicity adjustments in clinical trials is a very controversial issue. The aim of this paper is to provide general recommendations that can help determine whether multiplicity adjustment is necessary.
A brief overview of the most commonly encountered multiplicity adjustment methods and of the multiplicity issues that often arise in clinical trial settings (multiple endpoints, subgroups, multiple doses or treatments and interim monitoring) is given.
To provide an insight into the Italian context, protocols which have been evaluated by the Ethical Committee of the Policlinico of Padova from January 2009 to May 2010 are reviewed and discussed.
The literature suggests that there is no one single method for multiplicity adjustment that is preferred in all instances. 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.
Multiplicity adjustments are related to the need of the clinical research to fully exploit the information from clinical trials and their usage should be motivated by the effective benefits of exploiting such information at the maximum extent.

KEY WORDS: *multiplicity adjustment; clinical trials; adaptive designs; Bonferroni*

Introduction

Over the past decade, the multiplicity issue has been of increasing concern in the clinical trials area both from a statistical and regulatory perspective. Multiplicity arises in virtually any clinical trial because of multiple tests or looks upon data, e.g., for several treatment or dose groups, multiple endpoints, interim analyses, multiple subgroups, and at the very end from the need of the investigators to exploit the study answering as many questions as possible, keeping study integrity and correctness.
Multiple tests and looks of different aspects of a trial virtually widen the noise spectrum leading to an increased frequency of false signals (1). The fact that carrying out multiple statistical tests in a sin-

** A preliminary version of this paper has been presented at the II National Congress of BIAS (Biometristi dell'Industria Associati), Rome, 26th-28th of May 2010.*

gle experiment inflates the probability of a false-positive result (2-5), is well recognized by the medical research community, and this has promoted the development of many multiple comparison adjustment procedures.

Confirmatory clinical trials in drug development and approval are largely driven by expectations of the regulatory agencies. In 2002, the European Agency for the Evaluation of Medicinal Products (now European Medicines Agency, EMA, http://www.ema.europa.eu/) published a document concerning the multiplicity problem [6] stating that "…multiplicity can have a substantial influence on the rate of false positive conclusions whenever there is an opportunity to choose the most favourable results from two or more analyses…". Later, the EMEA reflection papers (7-8) on adaptive designs refer to the recommendations given in the previous document (6) for stating details of the multiplicity issues arising in the analysis plan.

At present, the U.S. Food and Drug Administration (FDA, http://www.fda.gov/), is drafting a guidance document (9) for industry on adaptive designs that echoes EMEA guidelines by defining the bias associated with the multiplicity of options as the bias introduced "…because of the opportunity to choose the successful result from among the multiplicity of options".

Based on concerns about elevated false-positive (Type I) error rate, not only regulatory agencies, but also scientific guidelines such as the CONSORT Statement (10) advocate the use of multiplicity adjustments addressing multiplicity of analyses as a possible limitation (Item no. 20), and suggest indicating pre-specified or exploratory ancillary analyses (Item no. 18 and 20 of CONSORT extension) (11).

While these documents acknowledge the non-optimal performance of multiplicity adjustment, provided that it is well documented and justified in the study protocol, this remains a very controversial point. On the opposite side of the debate is the recent EMA provocative statement, open for discussion, at the ISPE Meeting (http://www.pharmacoepi.org/meetings), stating that "... results from epidemiology and pharmacoepidemiology studies mislead the public because of a failure to adjust for multiple comparisons. Journals should not publish studies that do not account for multiplicity".

Many authors argue against multiplicity control for several reasons, just to name a few, to avoid the loss of statistical power (i.e., the probability of rejecting the null hypothesis given that the treatment is actually effective) when information from other sources should be used to prevent misinterpretation (12) or when there is a priori interest in estimating marginal treatment effects for a limited number of comparisons (13); to preserve organization and logistical efficiency in multi-arm trials (14); to prevent ad hoc corrections to manipulate statistical significance; and to avoid unrealistic global null hypothesis testing that no treatment has any effect and no subgroups benefit from the treatment (15). Others contend, however, that ignoring multiplicity can lead to serious misinterpretation of study findings and publishing bias (16).

The aim of this paper is to give general recommendations that can help determine whether multiplicity adjustment is necessary. The next section contains a brief overview of multiplicity adjustment methods. The following section introduces multiplicity issues that most commonly arise in clinical trial settings. The last section comprises some concluding remarks.

Overview of multiplicity adjustment methods

A large body of literature describes statistical methods to adjust Type I errors for multiple testing and we refer the reader elsewhere for the details [see, for example, the books by Hsu (17), Miller (18) and Hochberg and Tamhane (19)].

Until recently, most of the literature on multiple testing focused on methods to control the false-positive error rate for the entire trial, which is denoted as familywise error rate (FWER), where 'familywise' (also 'experimentwise') means for the selected hypotheses out of all possible hypotheses, at a given α level (that is, methods to ensure that the FWER $\leq \alpha$).

The most well-known method is the Bonferroni procedure, which sets the significance level for individual tests at divided by the total number of study comparisons. The Bonferroni method, however, yields conservative bounds on Type I error and, hence, it tends to reduce power, inflating the Type II error rate. When the total number of compari-

sons is large, the Bonferroni procedure is in addition overly conservative. Several modified and sometimes more powerful versions (20-22) of the Bonferroni method have been developed that provide strong control of the FWER.

To account for the hierarchical structure of multiple analyses, gatekeeping procedures that involve the prospective specification of families of null hypotheses that are tested in a sequential manner have been developed (23). The most straightforward application of these procedures is that to the multiple endpoint problem, via a sequential gatekeeping approach, in which testing of families (or gates) of null hypotheses in a pre-specified 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, 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. Different approaches to control the FWER can be used for this adjustment, including some of those discussed above. In other words, if the primary endpoint fails to achieve statistical significance, the study is regarded as a failure and no secondary endpoint will be tested. When the success or failure of the study is driven by the primary endpoint, this sequential strategy makes logical sense. In case of multiple doses or composite endpoints this approach is inappropriate (24).

The increase of the Type I error when repeated analyses are performed on accumulating data led several authors in the late 1970s to propose simple strategies, named Group Sequential Designs (GSDs), to maintain this risk of Type I error to a desired value, given a fixed schedule of analyses planned in advance, the so-called interim analyses, (see Jennison and Turnbull (25) for a review). The earliest proposals for GSDs developed by Pocock and O'Brien and Fleming, were primarily two-sided tests for normal responses with known variance. More flexible designs based on error spending functions which allow unequal group sizes and unspecified inspection times have also been developed. However, the GSD allows only reassessing the number of people to be involved in the mean course of the study, but no insight is given to the appropriateness of the chosen, expected, clinical effect, which remains a priori fixed as in the general, classical designs.

An alternative approach to the design of sequential clinical trials that well accomplishes this task has been proposed by Bauer (26) and Bauer and Köhne (27). This method is called the adaptive design approach (28). It allows a wide range of modifications to the trial design, including sample size re-estimation and design parameters re-specification, to be made at each interim analysis, while maintaining control of the overall Type I error rate (29).

Multiplicity issues

Multiple primary endpoints. 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, power, and other features of the clinical trial design will be based on the primary endpoint. In a regulatory context, when there is a single pre-specified primary efficacy endpoint and all additional endpoints are declared as providing only supportive exploratory information, adjustment for multiplicity will typically not be necessary.

Human diseases are often characterized by multiple measures, therefore the effect of the intervention in a clinical trial may actually be measured via multiple endpoints.

When significant results are required for all the primary endpoints, in this case called co-primary endpoints, no adjustment for multiplicity is necessary. Noteworthy, the overall power will be less than the smallest power for testing each co-primary endpoint. If the test statistics are completely independent, then the overall power is the product of the powers for testing each co-primary endpoint. If the test statistics are perfectly correlated, then the power for detecting the same (standardized) effect size on all endpoints is the same as the common power for detecting the same (standardized) effect size at the individual co-primary endpoint level (30). This should be taken into account at the sample size estimation step and explicitly documented in the study protocol.

Table 1. Need for multiplicity adjustments depending on the number of primary endpoints (p), the number of secondary endpoints (s) and the endpoints for which a significant result is required.

Primary Endpoints	Secondary Endpoints	Target (significance)	Multiplicity adjustment
$p = 1$	$s \geq 1$	One primary	No
$p = 1$	$s \geq 1$	One primary, the secondary marginally significant	Yes
$p \geq 1$	$s \geq 1$	All primary	No
$p \geq 1$	$s \geq 1$	Al least one primary	Yes
$p = 3$ (P1, P2, P3)	$s \geq 1$	Either P1 or both P2 and P3	Yes
$p = 3$ (P1, P2, P3)	$s \geq 1$	Either (P1 and P2) or (P1 and P3)	Yes
$p \geq 1$	$s \geq 1$	Hierarchical significance among primary	No
$p \geq 1$	$s \geq 1$	$k \leq p$ primary, the remaining marginally	Yes
$p \geq 1$	$s \geq 1$	Complex (usually hierarchy between primary and secondary)	Yes

When a significant result is required for only one of multiple primary endpoints in order to consider a trial positive, each endpoint must be tested with a significance level that has been corrected for multiplicity. There are also other circumstances in which multiplicity adjustment is usually avoidable, for example, when additional endpoints are used only to explore treatment mechanisms, to examine secondary hypotheses, or to generate hypotheses for future studies. Table 1 summarizes our subjective view of the need for multiplicity adjustment for various scenarios.
In order to distinguish between endpoints that could be used for labelling claims from endpoints that would provide only supportive evidence, some authors (31) advocate to clearly specify the set of endpoints for which strong control of the Type I error rate is guaranteed, and that any significant result within this set is equally valid, regardless of their designation (i.e. primary, secondary,…).

Subgroups.
A subgroup within a clinical trial may be classified as proper or improper depending on whether it is delineated by patients' baseline characteristics or by a post-randomization event or measure (32). No validity is usually given for improper outcome-based subgroup findings since a subgroup effect may be the result of inherent patients' characteristics that led to a particular response or side effect rather than a true treatment effect. Subgroup analyses have often been criticized for being post hoc and leading to suboptimal clinical practice (33). Nevertheless, there are many situations where there are prior clinical or biological reasons why a certain subgroup may particularly benefit from a given treatment. In this context, testing strategies for both the overall and pre-specified subgroup hypotheses, which have optimal power and strongly control the FWER exist (34) and are recommended. Strategies have been proposed that allow testing for a subgroup although results in the overall population might not be significant (35). If there is no a priori reason to expect subgroup differences, one should first test whether there is an effect in the overall group, and proceed further only if this test were significant. In this case, an interaction test (36) could be applied.

Multiple dose groups or treatments.
Identifying the minimum effective dose that produces a relevant biological effect is a fundamental step in the development of any medicinal drug. The analysis of data from dose-response studies has traditionally been divided according to two major approaches: model-based approaches and multiple comparison procedures. Model-based strategies assume a functional relationship between the response and the dose, taken as a quantitative factor, according to a pre-specified parametric model. On the contrary, multiple comparison procedures consider the dose as a qualitative factor and make very few or no assumptions about the underlying dose-response model. Unified strategies combining multiple comparison and modelling techniques while preserving the FWER in dose-response studies have been proposed (37). In randomized dose-finding studies it is common to compare several dose groups with a placebo

group and/or an active control. It is known that establishing equivalence to an active control is of limited value if both the dose and the active control have not been shown to be superior to placebo. Furthermore, it is well recognized that simultaneous testing for non-inferiority and superiority in an active-controlled clinical trial does not require multiplicity adjustments (38).

Since confirmatory clinical trials are an expensive and time-consuming component of drug development, multi-arm clinical trials in which multiple treatments may be tested simultaneously (14) are attractive for efficiency and logistical reasons. Some authors (14, 39) suggest that the need for multiplicity adjustments depends on the relatedness of individual comparisons. Only if each experimental arm is a component of a primary overall question (e.g. the treatment recommendation will be based on a joint interpretation of treatment comparisons), a multiplicity adjustment is required.

Interim Monitoring.

Interim assessments of long-running confirmatory clinical trials are important from both ethical and economic standpoints as in most medical studies and need to be done carefully and judiciously. The decision to conduct an interim analysis should be based on sound scientific reasoning that is guided by clinical and statistical integrity, standard operating practices, and regulatory concerns. Since interim analyses allow early stopping for strong evidence of efficacy or futility, there is potential for gaining for single entities such as pharmaceutical industries. In this context, multiplicity adjustments are extremely important and both GSDs and adaptive designs can be applied.

Adaptive designs should not be seen as a means to alleviate the burden of rigorous planning of clinical trials. Instead, adaptive designs would be best utilized to cope with difficult experimental situations. In all instances the type of the anticipated design modification would need to be described and justified in the study protocol.

Some insight in the Italian Experience

A review of the protocols which have been evaluated by the Ethical Committee of the Policlinico of Padova reveals (Table 2) that a small fraction of them is actually explicitly referring to the issue of multiplicity in taking multiple compari-

Table 2. Protocols evaluated by the Ethical Committee of Policlinico of Padova.

	Protocols	Drug studies	Controlled	Multi-group comparison	Adjustment in analysis
2009					
January	17	10	6	3	3
February	11	4	2	0	
March	19	11	9	2	2
April	18	10	8	2	2
May	25	14	10	4	4
June	15	12	6	1	1
July	23	9	6	2	2
September	26	18	11	3	3
October	20	15	10	1	0
November	16	10	7	1	1
December	16	4	1	0	
Total 2009	206	117	76	19 (9,2%)	18
2010					
January	17	6	4	3 (1 Bayes)	3
February	19	7	4	0	
March	21	7	4	1	1
April	31	14	12	3	3
May	15	3	2	0	

sons (9.2% in 2009 and 6.8% in 2010). The importance given to the proper analysis and definition of primary and secondary endpoints in the trials' protocol, which includes the treatment of multiplicity, is ranking between the 4th and the 6th position in terms of importance, as emerging from the NEBICE survey among biostatisticians operating in the Italian ethical committee (www.sismec.info). This seems not to be affected by the post-graduate level of education attained (Table 3).

Conclusions

Multiplicity adjustments are unquestionably the consequence of the need of the clinical research to exploit the results eventually obtained in clinical trials: this encompasses the need to reduce the lag between drug ideation, approval and its availability on the market, but also the ethically sensitive issue of getting as much information as possible from each single patient involved in the study.

In this sense, the usage of sequential and adaptive designs is perhaps one of the most promising perspective in clinical research over the last decades. This of course poses several issues, in terms of an ethically sound conduction of the study, but should not prevent it, in principle, from increasing their concrete application in nowadays trails.

The new flexibility introduced by adaptive designs to clinical trials entails a possibly high danger of malpractice. To prevent any fraudulent use, specific regulatory guidance is required on the prerequisites of any contingent modification of the original study plan.

Some specific points could be considered to reduce the negative aspects of the multiplicity adjustments still preserving their appealing characteristics of increasing study rationality:

1. Multiplicity should never be the effect of a poorly designed experiment. Its usage should be motivated by the effective benefits of exploiting the study information at the maximum extent.
2. Tests on primary and secondary endpoints, the hierarchy among them should be always clearly indicated in each protocol design, to allow regulatory agencies and ethical committee a proper evaluation of the risk of type I errors.
3. Power should be ensured and clearly calculated and discussed with reference to the actual set of hypotheses as implemented with multiplicity adjustment.
4. The loss of statistical power ensuing from multiplicity control or adjustment should be anticipated, and adequately foreseen, in sample size calculations.
5. External and independent ad-hoc steering committee should be ensured to evaluate proper study adaptation and its compliance with what indicated in the study protocol.

Regarding the methods suitable for adjustment, the literature suggests that there is no one single method for multiplicity adjustment that is preferred in all instances. Rather, the appropriate measure will depend on the study design, the primary research questions that are to be addressed, and the strength of inferences that are required. Regardless

Table 3. Ranks of importance of primary and secondary endpoint definition and analysis in evaluating protocols as resulting from the NEBICE survey.

Rank (%)	1	2	3	4	5	6	7	8	9	10
Overall										
Primary	2	0	10	11	18	24	18	9	4	4
Secondary	3	1	3	9	15	15	31	9	13	2
MD or MA										
Primary	0	0	5	22	22	10	18	2	2	8
Secondary	0	2	5	2	22	12	34	7	15	0
PhD										
Primary	3	0	12	6	16	26	18	11	4	3
Secondary	4	0	2	11	11	16	30	9	12	3

of the approach is used, the selected procedure must be specified in the clinical trial protocol and statistical plan before undertaking any analyses of the data. A posteriori procedures for the adjustment for multiple comparisons may appear as stopgap measures to avoid negative results and should therefore be avoided, even in exploratory analyses.

References

1. Huque MF, Al-Osh M, Dubey SD. P-Values, Evidence and Multiplicity Considerations for Controlled Clinical Trials. Encyclopedia of Biopharmaceutical Statistics. 2003:696-706.
2. Bauer P. Multiple testing in clinical trials. Stat Med. Jun 1991;10(6):871-889; discussion 889-890.
3. Godfrey K. Statistics in practice. Comparing the means of several groups. N Engl J Med. Dec 5 1985;313(23): 1450-1456.
4. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet. May 7-13 2005;365(9471):1657-1661.
5. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet. Apr 30-May 6 2005;365(9470):1591-1595.
6. Committee for Proprietary Medicinal Products (CPMP). Points to consider on multiplicity issues in clinical trials. 2002; http://www.ema.europa.eu/pdfs/human/ewp/090899en.pdf, 2010.
7. Committee for Medicinal Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. 2006; http://www.ema.europa.eu/pdfs/human/ewp/245902en.pdf, 2010.
8. Committee for Medicinal Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. 2007; http://www.ema.europa.eu/pdfs/human/ewp/245902enadopted.pdf, 2010.
9. Centre for Drug Evaluation and Research (CDER), Centre for Biologics Evaluation and Research (CBER). Adaptive Design Clinical Trials for Drugs and Biologics. Guidance for Industry. 2010; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf, 2010.
10. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. Apr 17 2001;134(8):663-694.
11. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. Mar 8 2006;295(10):1152-1160.
12. Saville DJ. Multiple Comparison Procedures: The Practical Solution. The American Statistician. 1990;44: 174-180.
13. Cook R, Farewell V. Multiplicity Consideration in the Design and Analisys of Clinical Trials. Journal of the Royal Statistical Society, Series A. 1996;159:93-110.
14. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. Clin Cancer Res. Jul 15 2008;14(14):4368-4371.
15. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. Jan 1990;1(1):43-46.
16. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. Apr 18 1998;316(7139):1236-1238.
17. Hsu JC. Multiple Comparisons: Theory and Methods. London: Chapman and Hall; 1996.
18. Miller RG. Simultaneous Statistical Inference. New York: Springer-Verlag; 1981.
19. Hochberg Y, Tamhane AC. Multiple Comparison Procedure. New York: John Wiley & Sons; 1987.
20. Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1988;75:800-802.
21. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979;6: 65-70.
22. Leon AC, Heo M, Teres JJ, Morikawa T. Statistical power of multiplicity adjustment strategies for correlated binary endpoints. Stat Med. Apr 15 2007;26(8): 1712-1723.
23. Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Stat Med. Aug 15 2003;22(15):2387-2400.
24. Hung HM, Wang SJ. Some controversial multiple testing problems in regulatory applications. J Biopharm Stat. 2009;19(1):1-11; discussion 12-41.
25. Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. New York: Chapman & Hall/CRC; 2000.
26. Bauer P. The choice of sequential boundaries based on the concept of power spending. Biometrie und Informatik in Medizin und Biologie. 1992;20:130-148.
27. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses Biometrics. 1994;51:1029-1041.
28. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development—an Executive Summary of the PhRMA Working Group. J Biopharm Stat. May 2006;16(3): 275-283; discussion 285-291, 293-278, 311-272.
29. Chow SC, Chang M. Adaptive design methods in clinical trials - a review. Orphanet J Rare Dis. 2008;3:11.
30. Offen W, Chuang-Stein C, Dmitrienko A, et al. Multiple Co-primary Endpoints: Medical and Statistical Solutions. A Report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America. Drug Information Journal. 2007;41: 31-46.
31. Snapinn S, Jiang Q. Analysis of multiple endpoints in clinical trials: it's time for the designations of primary, secondary and tertiary to go. Pharm Stat. Dec 22 2009.
32. Hirji KF, Fagerland MW. Outcome based subgroup

analysis: a neglected concern. Trials. 2009;10:33.
33. Lagakos SW. The challenge of subgroup analyses—reporting without distorting. N Engl J Med. Apr 20 2006;354(16):1667-1669.
34. Song Y, Chi GY. A method for testing a prespecified subgroup in clinical trials. Stat Med. Aug 30 2007;26(19):3535-3549.
35. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. Stat Med. Jan 15 2009;28(1):3-23.
36. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. Oct 15 2002;21(19): 2917-2930.
37. Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics. Sep 2005;61(3):738-748.
38. Tsong Y, Zhang J. Simultaneous test for superiority and noninferiority hypotheses in active-controlled clinical trials. J Biopharm Stat. 2007;17(2):247-257.
39. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials. Dec 2000;21(6):527-539.