

GUEST EDITOR'S NOTE: A EUROPEAN CONCEPT FOR GOOD STATISTICAL PRACTICES IN GLOBAL DRUG DEVELOPMENT

ANDREAS ZIPFEL

Section Biometric & Data Management, Department des Enregistrements et de L'Information Scientifique,
Synthelabo Recherche, Bagneux Cedex, France, and Program Co-chairperson

THIS NOTE INTRODUCES discussion papers from the DIA Workshop "European Concept for Good Statistical Practice in Global Drug Development," April 27-28, 1994, Edinburgh, Scotland, United Kingdom.

Since its release, the draft version of the "Note for Guidance on Biostatistical Methodology in Clinical Trials in Drug License Applications" (111/3630/92 draft 4) of the Committee for Proprietary Medicinal Products (CPMP) Working Party on Efficacy of Medicinal Products has been widely discussed and many position papers are circulating among expert groups at a national and international level from industry, academia, and regulatory agencies. This document will have a great impact on European regulatory practice. For this reason, there was an urgent need for a general European concept for good statistical practice (GSP) comprising the views of professionals with different backgrounds (industry, academia, and

regulatory agencies), and from the different reReprint address: Andreas Zipfel, Section Biometric

& Data Management, Departement des Enregistrements et de L'Information Scientifique, Synthelabo
Recherche, 31 avenue Paul Valiant Couturier, BP
110, 9225 Bagneux Cedex, France.

vant disciplines, such as statistics, data management, clinical research, and regulatory affairs. A European concept must also take into account the Food and Drug Administration (FDA) standards and other international regulatory requirements, since most pharmaceutical companies operate on a global basis.

In order to promote a general European GSP concept, it was felt that a bridge between statisticians working in an industrial/clinical environment and those who are evaluating clinical/statistical applications in regulatory agencies needed to be built. This will help in better understanding the needs and the different perspectives in evaluating the benefits and risks of new medical compounds.

Obviously, a single workshop cannot establish a European GSP concept, but it is felt that the background papers and the discussions revealed many important aspects in how to deal with statistical matters in new drug applications. It became clear that a consensus is possible.

The program chairpersons would like to thank the panel members and the working groups from academia, regulatory agencies, and industry for their personal commitment and the highly competent contributions, and the DIA staff for their professional and encouraging help.

471

Drug Information Journal, Vol. 29, p. 473, 1995 Printed in the USA. All rights reserved.
0092-8615/95

Copyright © 1995 Drug Information Association Inc.

DISCUSSION PAPERS: A EUROPEAN CONCEPT FOR GOOD STATISTICAL PRACTICE IN GLOBAL DRUG DEVELOPMENT

PROGRAM Co-CHAIRPERSONS

ANDREAS ZLPFEL

EFSPI President, Synthelabo Recherche, France

ANNETRE ROBERTSON

EFSPI Council, Zeneca Pharmaceuticals, United Kingdom

MARCO GIRELLI
EFSPI Council, Glaxo SpA, Italy

KARSTEN SCHMIDT
EFSPI Council, Spadille Biostatistik Aps, Denmark

The following papers are from the DIA Workshop "A European Concept for Good Statistical Practice in Global Drug Development," April 27—28, 1994, Edinburgh, Scotland, United Kingdom. These papers include background information and discussion on topics related to the statistical contents of single reports and statistical procedures in a new drug application.

Reprint address: Adreas Zipfel, Synthelabo Recherche, 31 Avenue Paul Vaillant Couturier, BP 110, 92225 Bagneux Cedex, France.

473

Drug Information Journal, Vol. 29, pp. 475-477, 1995 Printed in the USA. All rights reserved.
0092-8615/95
Copyright ~ 1995 Drug Information Association Inc.

**HOW MANY POPULATIONS MUST BE
ANALYZED AND HOW SHOULD THEY
BE DEFINED (INTENTION-TO-TREAT,
ELIGIBLE, PER PROTOCOL
POPULATION, ETC.)?**

DISCUSSION COORDINATOR: KLAS SVENSSON

Astra Draco AB, Sweden

THE COMMITTEE FOR Proprietary Medicinal Products (CPMP) “Draft Guidelines on Biostatistical Methodology” state in Part 9.2 that: “In the protocol of most trials an intention-to-treat (ITT) population should be defined.” The guidelines also state that the definition of this concept depends on the design of the study and must be defined. Potentially acceptable definitions are suggested: all randomized patients in randomized groups; all randomized patients who have at least one evaluation after baseline; all randomized patients who are correctly allocated and have the disease which is under study; all randomized patients who are correctly allocated, have the disease under study, and had received at least one dose of the drug.

The Food and Drug Administration (FDA) “Guidelines for the Format and Content of the Clinical and Statistical Sections of New Drug Applications” state in Section III.B.9.a that: “As a general rule, even if the applicant’s preferred analysis is based on a reduced subset of the patients with data, there should be an additional intent-to-treat’ analysis using all randomized patients.” Section III.B .9.c.2. b states that: “The results of a clinical trial should be assessed not only for the subset of patients who completed the study, but also for the entire patient population randomized (the intent-to-treat analysis).” The draft EC guidelines’ phrasing reflects the experience in application of the concept

475

since the FDA guidelines were published; ITT is not an obvious concept, and must be defined in each case, depending on the circumstances.

It should be stressed that ITT is a concept rather than a method. Behind this concept lies the problem with missing data; thus, the handling of dropouts and withdrawals must be addressed in an integrated way. The last value carried forward (LVCF) method, a common method applied in this situation, is a simple and easy way to deal with the problem, and the method seems to be accepted by regulatory bodies.

The per-protocol approach (this term is preferred instead of “valid cases” or “efficacy population”) is defined as eligible patients with no major protocol violations. Before breaking the blind of a trial, all protocol violations should be classified as major or minor, where minor implies no vital importance for the outcome of the study.

The ITT approach is normally the one preferred in the later Phases III and IV, when efficacy is the main interest and when the study is a management rather than an explanatory one. Studies of this kind are often large, multicenter, multinational studies with outpatients; the possibility of checking protocol violations such as time and amount of study medication intake are usually limited. In early phases, where the objective of a study is pharma

:1

476

Discussion Coordinators

cokinetic or pharmacodynamic, it is more suitable to apply a per-protocol approach.

The ITT approach interpreted as “all randomized patients” can be suitable when randomization has been performed in an early stage of the study, and withdrawal may be dependent on the intended treatment. The “all randomized patients” approach is also preferred by those who advocate that the randomization principle is the basis for the statistical inference of a study; any violation of the randomization procedure will make the results more or less invalid. For a more clinical point of view, it is argued that patients who did not take any study medication should not be included in the evaluation of the study.

Similar arguments can be used when discussing the noneligible or nonqualified patients in a study, that is, patients who violate the inclusion/exclusion criteria, do not have the intended disease, or do not receive any study medication.

For safety evaluation, an ITT approach interpreted as “at least one dose” is often used. It can, however, be argued that this may lead to a diluted effect when the rate or incidence of adverse events must be assessed. Thus, an approach with patients who have been on

treatment for a certain time can be more reasonable.

For bioequivalence studies, it seems more reasonable to adopt a per-protocol approach, as the primary question is pharmacodynamic.

Specific questions discussed were:

1. In drug trials, where the comparison is either placebo or active drug treatment, should the ITT concept be interpreted as “all patients treated” – all randomized patients who have received at least one dose of the study medication?
2. Could a presentation of both approaches, at least for the primary variable(s), facilitate the interpretation of the study and be a starting point for judging the robustness of the conclusions of the study?
3. Is the principle of exclusion of nonqualified patients acceptable for an ITT evaluation?
4. Can carrying forward baseline measurements to treatment period assessments be justified?
5. Is there a need and/or justification for an additional intermediate analysis between a “puristic” ITT and a “squeaky clean” per-protocol? Does such an approach strengthen the findings in the study?

This position paper has been prepared by a working group within FMS, the Swedish Association for Medical Statistics:

Hans Wedel, the Nordic School of Public Health; Mikael Aström, Pharmacia; and Klas Svensson, Astra Draco AB.

DISCUSSION SUMMARY

All members of the panel repeatedly stressed during the discussion that the concept of ITT, as it is understood today, is a principle that is normally the preferred approach for confirmatory studies in phase LIIA. The concept is only a principle, however, and the implementation of this principle may vary from case to case, depending on the type of study, treatment, and so forth. “Include as many as you can” and “use your judgment for implementation” were statements made during the discussion. Once a patient has been randomized into the trial, every effort should be made to assess the patient and include him/her in the evaluation of the trial. In addition to the randomization principle as a motivation for the ITT approach, the recent request from FDA for a broader patient population in later phases was mentioned.

There are cases where patients could be excluded from evaluation: women randomized into a trial for men only was mentioned as an artificial example, whereas patients whose blood pressure is not high enough to render them eligible for a trial on high blood pressure should be retained for the evaluation. For patients not treat-

Discussion Papers: A European Concept for Good Statistical Practice

ed at all, a principle could be to exclude them if one is sure that the exclusion was unrelated to treatment, and include them otherwise. For safety evaluation, the interpretation of the concept as “at least one dose of study drug taken” seemed to be generally accepted.

Requests from industry statisticians in the audience for more details in regulatory guidelines specifying how and when to use ITT were turned down by panel members from the regulatory authorities, who pointed out that as there seemed to be a consensus about the principle, there was no need for more details. This standpoint was supported by industry statisticians in the audience: "A good statistical education should provide the judgment needed for implementation of the principle." A certain implementation can be defended by a sensitivity analysis using other choices, and showing the robustness of the evaluation of the trial.

Certain instances, among them crossover trials and bioequivalence trials, were mentioned, where the ITT principle seemed to be unsuitable as a first line approach. Comparative trials with an active control

may also be an instance where ITT is unsuitable in certain cases.

The problem of replacing missing data is closely connected with the ITT principle. The last value carried forward method was referred to in the discussion as an accepted method, provided it is applied in a conservative way, not favoring the test drug. It was suggested that if the frequency of missing data exceeds two or three percent of all data, a comparison between "worst case scenario" and "best case scenario" should be performed to strengthen the conclusions from the trial.

CONCLUDING REMARKS

A member of the audience remarked at the end of the discussion that there was obviously a great consensus in the application of the ITT principle, both from the regulatory and the industry sides. The concept, however, seemed to trigger a number of questions about the implementation of the principle, which requires that statisticians rely on experience and good statistical practice rather than more detail in regulatory guidelines.

477

HOW MUCH DETAIL ON CONFIRMATORY STATISTICS AND EXPLORATORY STATISTICS MUST A STATISTICAL REPORT CONTAIN?

DISCUSSION COORDINATORS

ADREAS ZIPFEL

Synthelabo Recherche, France

PAUL GROB

Mirai, The Netherlands

THE STATISTICAL methodology paragraph of the protocol is logically deduced from the study objectives and the intentions that justified the trial. The clinical results (the data), however, reflect what has actually been *obtained*.

The confrontation of the expected with the observed lies at the heart of statistical investigations. For correct inferential statistics it is crucial to know whether the hypotheses were clearly formulated *a priori*, that is, in the protocol, or not.

In order to obtain scientifically sound results and to avoid inflating the global alpha error risk, the study objectives and hence the primary outcome measures must be limited to a small number, not more than two or three.

There is frequently a considerable time lag, however, between the initiation of a study and the availability of the study results. During this time many events may intervene:

- Results of other studies may modify the assumptions that were made at the planning stage or may even respond in an unexpected way to some of the objectives and study conditions,
- The actual acquisition of the results may deviate from the procedures as described in the study protocol (protocol violations, difficulties in the recruitment of patients, unbalanced distribution of sample size between study centers, etc.),
- Occurrence of baseline differences despite randomization, or
- The development of new or improvement of old statistical procedures that may be more appropriate or more powerful for the analysis of the data described in the protocol and collected in the study.

Not all events possible during the actual course of a study can be forecasted; some may bring the study assumptions into question.

The clinician responsible for the study, together with the statistician, usually investigate clinical data much closer than strictly necessary in the inferential sense. Frequently, regulatory authorities ask for supplemental *a posteriori* analyses (Europe) or perform them themselves (United States). By performing *sensitivity analyses* they want to assure that the results are sufficiently robust.

If, on the other hand, the sponsoring company takes the initiative in presenting additional *a posteriori* results, this may easily be criticized in the review process as invalid.

Questions still remain however:

480

Discussion Coordinators

- How many objectives can be supported by a single clinical trial? What kind of multiple analyses and endpoints are necessary?
- Should, in all studies, routine statistical testing be performed, checking for gender effects, baseline differences, and so forth? What should be done if some of these “tests” turn out to be “statistically significant”?
- To what extent can exploratory findings invalidate explanatory results?
- How much of the “expanded” investigation performed by the sponsor must be reported in the context of a drug license application?
- How many sensitivity analyses can be reasonably asked for and how could their impact be anticipated in the presentation of statistical results to registration authorities?

SUMMARY OF THE DISCUSSION

The first discussion topic was routine statistical testing of baseline factors. The discussants agreed

with Gary Koch's distinction of two possible interpretations of small p-values; the p-value leading to a contradiction/rejection of the null hypothesis in the context of statistical testing or the p-value describing rare events.

In connection with baseline descriptions, it was felt inadequate to interpret small p-values as statistically significant. It was suggested that p-values should be made available to reviewers as part of descriptive statistics and it should be left to them to decide whether to dismiss the p-values or not. This procedure has the advantage of being completely transparent and the sponsor cannot be suspected of trying to hide problems that occurred during the randomization.

If small p-values in connection with baseline factors indicate imbalances between treatment groups, adjusted analyses can elucidate how sensitive the conclusions are to imperfection in the randomization.

These "data-driven" adjusted analyses cannot replace the primary analysis. Study sponsors and reviewers, however, should be particularly sensitive in a double hit situation (Gary Koch), that is, when confronted simultaneously with small sample sizes *and* high correlations between baseline factors and study outcome (prognostic factors).

A clear preference was given to the case where key baseline factors that may influence the study conclusions had already been identified in the statistical methodology paragraph of the protocol and incorporated into the statistical model. Studies are not usually adequately designed to detect baseline differences and even apparently "statistically insignificant" imbalances can be clinically pertinent, especially in the above mentioned double hit situation. This observation, together with the problems raised by the multiplicity issue, are reasons why routine mechanisms such as data-driven adjusted analyses would not be an adequate solution.

If potential prognostic factors were not sufficiently dealt with in the study protocol, it was suggested that these factors could still be identified through a triple blind analysis in preparation of a report planning meeting where the statistical methodology could still be updated if necessary, prior to breaking the randomization code. This option is already part of the standard operating procedures in some companies. It should be mentioned, however, that this option too is a data-driven decision but with the advantage that it prevents deliberate introduction of bias in favor of one of the treatments into the analysis. The question was left open as to whether a partial decoding of groups (without attributing the treatments to the groups) is acceptable as a general consensus.

The second discussion topic dealt with the problem of subgroup analyses. In order to stimulate the discussion, the result of the subgroup analysis by astrological birth sign in the 1515-1 trial of 16,000 patients with suspected acute myocardial infarction was presented (1). The results



- How many objectives can be supported by a single clinical trial? What kind of multiple analyses and endpoints are necessary?
- Should, in all studies, routine statistical testing be performed, checking for gender effects, baseline differences, and so forth? What should be done if some of these "tests" turn out to be "statistically significant"?
- To what extent can exploratory findings invalidate explanatory results?
- How much of the "expanded" investigation performed by the sponsor must be reported in the context of a drug license application?
- How many sensitivity analyses can be reasonably asked for and how could their impact be anticipated in the presentation of statistical results to registration authorities?

SUMMARY OF THE DISCUSSION

The first discussion topic was routine statistical testing of baseline factors. The discussants agreed with Gary Koch's distinction of two possible interpretations of small p-values; the p-value leading to a contradiction/rejection of the null hypothesis in the context of statistical testing or the p-value describing rare events.

In connection with baseline descriptions, it was felt inadequate to interpret small p-values as statistically significant. It was suggested that p-values should be made available to reviewers as part of descriptive statistics and it should be left to them to decide whether to dismiss the p-values or not. This procedure has the advantage of being completely transparent and the sponsor cannot be suspected of trying to hide problems that occurred during the randomization.

If small p-values in connection with baseline factors indicate imbalances between treatment groups, adjusted analyses can elucidate how sensitive the conclusions are to imperfection in the randomization.

These "data-driven" adjusted analyses cannot replace the primary analysis. Study sponsors and reviewers, however, should be particularly sensitive in a double hit situation (Gary Koch), that is, when confronted simultaneously with small sample sizes *and* high correlations between baseline factors and study outcome (prognostic factors).

A clear preference was given to the case where key baseline factors that may influence the study conclusions had already been identified in the statistical methodology paragraph of the protocol and incorporated into the statistical model. Studies are not usually adequately designed to detect baseline differences and even apparently "statistically insignificant" imbalances can be clinically pertinent, especially in the above mentioned double hit situation. This observation, together with the problems raised by the multiplicity issue, are reasons why routine mechanisms such as data-driven adjusted analyses would not be an adequate solution.

If potential prognostic factors were not sufficiently dealt with in the study protocol, it was suggested that these factors could still be identified through a triple blind analysis in preparation of a report planning meeting where the statistical methodology could still be updated if necessary, prior to breaking the randomization code. This option is already part of the standard operating procedures in some companies. It should be mentioned, however, that this option too is a data-driven decision but with the advantage that it prevents deliberate introduction of bias in favor of one of the treatments into the analysis. The question was left open as to whether a partial decoding of groups (without attributing the treatments to the groups) is acceptable as a general consensus.

The second discussion topic dealt with the problem of subgroup analyses. In order to stimulate the discussion, the result of the subgroup analysis by astrological birth sign in the 1515-1 trial of 16,000 patients with suspected acute myocardial infarction was presented (1). The results

1

Discussion Papers: A European Concept for Good Statistical Practice

showed an overall percentage reduction in odds of death during treatment period of 15% ($p < 0.05$); this percentage reduction was 48% ($p < 0.04$) in the subgroup of patients with astrological birth signs of Scorpio and 12% (not significant) for all others. Quoting the comments of the authors: "Even though the effect of treatment appears to be greater for

those born under Scorpio, it seems unlikely that the effects of treatment are really affected by birth sign!" The question was raised, however, as to how this result would have been interpreted if, by chance, this had been a subgroup with some clinical relevance determined by gender or age, for instance.

In the discussion that followed, an obvious consensus was very quickly established on the following points:

- The analysis of prespecified subgroups has more impact than data-driven subgroup analyses. There are some standard subgroups that should be routinely investigated: gender, age, body weight, impaired renal or liver functions, smokers! nonsmokers, and so forth; some of these subgroups are indication-specific, but all of them can be identified *a priori*, and
- It makes a difference as to the credibility of the results if subgroup analyses were required by regulatory authorities or presented as an additional option by the sponsor. Preferably, however, the requested subgroup analyses should be prespecified or at least an "early warning" mechanism should exist by which regulatory authorities may signal their concern. In this way, slowing down the review process by additional requests *a posteriori* may be avoided.

Major problems emerge if subgroup analyses are entirely data driven. It was considered very unlikely that a sponsor could successfully base additional claims (eg, widening the indication of the drug) only on data-driven subgroup analyses. There are situations, however, where this sort of investigation delivers valuable information, for instance, in the examination of unexpected (adverse) events.

Regulatory statisticians made it quite clear that when confronted with subgroup analyses, the leading questions for them are: "Why were subgroup analyses performed and why are they reported?" They expect concise answers to these questions.

Sponsors and reviewers should look at the entire package of a drug license application as a whole; in this context, subgroup analyses can make a good story if they give rise to subsequent confirmation by other studies or prove to be reproducible among studies and within the published literature.

It was mentioned that a distinction should be made as to whether there is an overall significant result or not. If not, subgroup analyses are purely explorative and hypotheses generating. In the presence of an overall significant result, however, subgroup analyses can be a valuable measure of the heterogeneity of the effect. In order to avoid too much reliance on chance findings, the consistency among studies must always be checked.

The philosophy of subgroup analyses as a tool of hypothesis generation, subsequently to be reconfirmed by studies specifically designed for this purpose, was judged as somewhat unrealistic in most cases due to the high costs in the conduct of clinical trials. Very often, the proof of whether data-driven exploratory results are valid or not can only be established after marketing authorization. It was also mentioned that statistical testing of subgroup factors is not always adequate and should more often be replaced by estimation of the magnitude of these effects, which need to be evaluated in terms of biological and clinical plausibility.

REFERENCE

1. Collins R, et al. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: The need for systematic overviews. *Stat Med.* 1987;6:245—250.

STATISTICAL TESTS AND ESTIMATIONS

Discussion COORDINATOR: KARSTEN SCHMIDT

Spadille Biostatistik ApS, Denmark

BACKGROUND PAPER

Introduction

IN GENERAL, THE topic is wide-ranging and must be restricted to a few fundamental aspects pertinent to trials on medicinal products forming the basis for license applications. For many years, medical research has overrated the importance of p-values and thereby statistical and clinical significance have been mixed up and misinterpreted. Among statisticians, there is consensus that confidence intervals (CIs) are much more informative, and that p-values give only marginally supplementary information. Nevertheless, it is apparently difficult to change the tradition, and the vast majority of superiority studies are still dimensioned on the basis of providing sufficient power for hypothesis testing. Also, $p < 0.05$ seems still to be the primary criterion for claiming superiority.

In equivalence studies, confidence interval approaches are more frequently used, and in bioequivalence studies, it is a regulatory requirement both in the United States and European Union to use “the 90% confidence interval wholly within the \sim [minimum relevant difference] MIREDLF range” as a criterion for equivalence (Figure 1). (MIREDLF is also called the smallest clinically meaningful difference.) It is not difficult to determine the sample size necessary to have a given power, that is, the probability of demonstrating equivalence by this criterion conditional on the appropriate standard deviation and expected true difference between treatments, since both tables and computer programs are available.

483

Sample size determination for superiority studies based on hypothesis testing has never been unambiguously defined and is, therefore, frequently abused. The textbook prescription is to dimension the study based solely on the MIREDLF (and a variability estimate, of course), that is, not influenced whatsoever by the expected difference for the investigational drug. This may lead to unnecessary human experimentation in cases where the investigational drug is much more efficient than that worthwhile achieving from a clinical point of view (the MIREDLF). Only by a group sequential design can unnecessary human experimentation be avoided in cases of an extreme effect. The traditional way of dimensioning a study assures that a clinically worthwhile effect is not overlooked but only in the sense that a high probability of rejecting the null hypothesis of no effect is achieved if prior assumptions are not violated. If, however, in case of rejection the CI overlaps the \sim MIREDLF range (Figure 1.b), there is no appropriate assurance that one of the treatments is superior, and this ought to be clearly stated in the report. Even the case of simultaneous statistical significance and equivalence is possible (Figure 1.c). Thus, the traditional approach to sample size determination does not assure a high probability of giving sufficient evidence of superiority by the CI method, on the contrary, if the true effect is the smallest clinically meaningful, it is almost certain that the confidence interval will not be wholly outside the MIREDLF range,

It is a natural consequence of recommending the use of CI that both types of studies should be dimensioned to have suf



----- One treatment superior

Inconclusive

Equivalence may be claimed (*) Equivalence may be claimed (NS)

+ +
O +MIREDIF

NS = Not significantly different from zero.

= Significantly different from zero.

FIGURE 1. Decision criteria based on confidence

intervals.

ficient expected power of leading to a conclusion of either superiority or equivalence, and that a claim of superiority is based on a CI wholly outside the MIREDIF and not on $p < 0.05$ only.

Consensus Statements to be Discussed

1. Claim of either superiority or equivalence should be based on the criteria of 90% CI wholly outside or wholly inside the MIREDIF range, respectively. If neither criterion is fulfilled, the study will be inconclusive,
2. A study should be dimensioned to have a sufficient expected power (normally $> 80\%$) of reaching a conclusion conditional on an appropriate variability estimate and prior expectation of the true treatment difference,
3. The protocol must: prespecify the primary variable(s) to make CIs for; give a thorough justification for the choice of MIREDIF, expected true treatment difference, and variability estimate; and give details about how power was calculated; and
4. Retrospective power calculations should be avoided since the CIs give the appropriate information of the ability of the study to have detected various true treatment differences.

Comments

Although some examples of inapplicable CIs exist, CIs may be determined in most practical situations where hypothesis testing is possible as the set of all simple null hypothesis that could be accepted, that is, the use of CIs does not seriously restrict the possibilities of making multiple comparisons, adjustment for sequential looks, and so forth. Where it is necessary to have more than one primary variable, confidence regions and MIREDIF regions can be used alternatively.

Any unbiased type of parametric or nonparametric CI can be chosen freely as long as prespecified either in the protocol or in the statistical analysis plan that has been finalized before breaking the blind. If a less efficient method is chosen, the sample size must be increased correspondingly. The statistical analysis plan, prepared on the basis of fully blinded data, must prespecify exactly which patients to include in the population forming the

a.

b.

c.

d.

+
-MIREDF

Difference

Discussion Papers: A European Concept for Good Statistical Practice

sis for the CI determination and how to handle missing observations. It will sometimes be appropriate to work on a logarithmic scale such as bioequivalence where this is frequently used.

Although the Bayesian method with other than a noninformative prior cannot, in general, be acceptable to regulators as a basis for approval of a new drug application, a subjective prior distribution of true treatment difference seems to be acceptable as a basis for sample size determination.

It is essential that the justification for the choice of MIREDF is also acceptable to the regulators. A poor choice of expected true treatment difference and variability estimate will certainly not be in the applicant's best interest since it will increase the costs of reaching a conclusion; nevertheless, it seems reasonable that the regulators require documentation in the protocol that the applicant has planned the study on a reasonably firm basis. The distinction between superiority and equivalence trials will largely disappear by consistent usage of the CI method.

As compared with the textbook prescription of traditional sample size determination for hypothesis testing, the CI method will result in smaller sample sizes if the true treatment difference substantially exceeds the MIREDF. Otherwise, the sample size will be larger, and in some cases the study will be unfeasible due to this. It seems to be an advantage in preventing probably inconclusive studies from being carried out at all.

The discussion coordinator, Karsten Schmidt, started by pointing out that he wanted to keep things simple and only focus on a very few fundamental concepts. Apparently, the simplicity had created some misunderstandings regarding the background paper, which needed clarification.

Four situations were sketched in Figure 1, all in the direction of a positive difference suggesting a beneficial effect of the test treatment as compared with the control treatment. Except for the possible choice of an asymptotic MIREDF range, however, the figure could as well be mirrored around zero and everything would still apply since case (a) states: "one treatment superior" and not the test treatment superior.

Even though for simplicity the figure presented the MIREDF ranges as \sim MIREDF, that is, symmetric around zero, the range could as well be asymmetric with, for example, the lower limit, in the direction of the test treatment being worse than the control treatment, closer to or even at zero.

Also, for simplicity, a difference in means was taken as an example. The principle could equally well be applied to any parameter, for example, the odds ratio with a MIREDF range around one or the log odds ratio.

Again, to keep things simple, Karsten Schmidt suggested not discussing how the minimum relevant difference range is to be defined even though this is an important issue. Hence, the basis for the discussion was that the MIREDF range is established one way or the other in a way agreed upon. Under this assumption the decision criterion based on CI in Figure 1 was proposed as consensus statement 1.

Stephen Senn opened the panel discussion by expressing the opinion that no confidence interval and no statistical technique gets around the particular difficulties which exist with equivalence studies in general. He gave some examples to illustrate the philosophical difficulties with equivalence versus difference and referred to a paper by Robert Temple (1). Therefore, he disagreed with consensus statement 1 covering equivalence and superiority trials within the same framework. Nonetheless, he recommended CIs as the way forward for reporting the results of clinical trials. Regarding consensus statement 4 about retrospective power calcula

485
U

tions, he was in 100% agreement with the proposal.

Stephen Senn mentioned that there are many different definitions of a clinically relevant difference, in practice, it is the difference one would not like to miss and not at all a difference that one claims would exist. In his opinion, the use of a clinically relevant difference as a means of distinguishing between drugs which are superior or equivalent would need yet another difference that another name must be found for.

Deborah Ashby said that retrospective power calculations tell her nothing. If used by applicants as a reason for confidence intervals that are too wide it could really irritate statisticians in regulatory authorities, demonstrating that the study was badly planned and should be enlarged to provide proper confidence intervals.

Joachim Röhmel said that although he had doubts about the existence then based on the assumption that the minimum clinically relevant difference exists, he agreed with some of the statements. His main message is that clinicians will not define what is really clinically relevant, and since even very small differences may be important, the limit should be zero.

Uwe Ferner heavily supports the use of confidence intervals but believes in p-values, too. Multiplicity problems and secondary objectives call for p-values to obtain some information on what to look for in further trials.

Sylvain Durrleman remarked that it seems as if the suggestion was to design the study based on an expected difference for the investigational drug. This would create difficulties in

defining what expectation to use since it could be the clinician's, the statistician's, the patient's, and so forth. It would be preferable to stick to the minimum difference that one would not like to miss because an overly optimistic expectation may lead to an underpowered study.

Gary Koch commented from the floor that although confidence intervals are very useful, they can be formulated in various ways such as difference of means, ratio of means, percent change, relative risks odd ratios, and so forth, and one must choose among them, whereas p-values are often robust to the metric, such as with categorical data.

Deborah Ashby responded to Gary Koch's comment that what she wants to see for a study is first of all, an estimate on a sensible scale and secondly, some estimate of how precise that estimate might be. Given the context, it is usually fairly well established what that scale should be. P-values come quite a long way down the list. It is disappointing how often that sort of sense failed to be in the summary of a study.

Willi Maurer responded to Stephen Senn that equivalence trials and the CI concept *are* being used, right or wrong and regardless of philosophical problems. The power attached to a difference expresses the importance attached to it. If the true difference is just above the minimum relevant difference, then by the 90% CI approach there is only 5% power attached, so probably the minimum relevant difference is very small so it does not matter if it is missed. Given this, he thinks one could use the concept but could as well take zero as the MIREDF and just go on the usual way.

John Lewis was not entirely happy with all the conclusions in the background paper, but he thinks that the consequences of adopting the suggested strategy would actually be beneficial. Generally, a study is regarded as inadequate if it just achieves the 5% level and has a confidence interval that just excludes zero. To convince doctors, sometimes mega studies are needed where the confidence interval lies miles away from zero and with a really small pvalue. To convince regulators, there is probably an element of that there too, and the two or more pivotal studies question which raises its head from time to time may be because actually $p = 0.05$ is not enough so people want to see two multiplied together and thereby have a smaller

tions, he was in 100% agreement with the proposal.

Stephen Senn mentioned that there are many different definitions of a clinically relevant difference, in practice, it is the difference one would not like to miss and not at all a difference that one claims would exist. In his opinion, the use of a clinically relevant difference as a means of distinguishing between drugs which are superior or equivalent would need yet another difference that another name must be found for.

Deborah Ashby said that retrospective power calculations tell her nothing. If used by applicants as a reason for confidence intervals that are too wide it could really irritate statisticians in regulatory authorities, demonstrating that the study was badly planned and should be enlarged to provide proper confidence intervals.

Joachim Röhmel said that although he had doubts about the existence then based on the assumption that the minimum clinically relevant difference exists, he agreed with some of the statements. His main message is that clinicians will not define what is really clinically relevant, and since even very small differences may be important, the limit should be zero.

Uwe Ferner heavily supports the use of confidence intervals but believes in pvalues, too. Multiplicity problems and secondary objectives call for p-values to obtain some information on what to look for in further trials.

Sylvain Durrleman remarked that it seems as if the suggestion was to design the study based on an expected difference for the investigational drug. This would create difficulties in defining what expectation to use since it could be the clinician's, the statistician's, the patient's, and so forth. It would be preferable to stick to the minimum difference that one would not like to miss because an overly optimistic expectation may lead to an

underpowered study.

Gary Koch commented from the floor that although confidence intervals are very useful, they can be formulated in various ways such as difference of means, ratio of means, percent change, relative risks odd ratios, and so forth, and one must choose among them, whereas p-values are often robust to the metric, such as with categorical data.

Deborah Ashby responded to Gary Koch's comment that what she wants to see for a study is first of all, an estimate on a sensible scale and secondly, some estimate of how precise that estimate might be. Given the context, it is usually fairly well established what that scale should be. P-values come quite a long way down the list. It is disappointing how often that sort of sense failed to be in the summary of a study.

Willi Maurer responded to Stephen Senn that equivalence trials and the CI concept *are* being used, right or wrong and regardless of philosophical problems. The power attached to a difference expresses the importance attached to it. If the true difference is just above the minimum relevant difference, then by the 90% CI approach there is only 5% power attached, so probably the minimum relevant difference is very small so it does not matter if it is missed. Given this, he thinks one could use the concept but could as well take zero as the MIREDF and just go on the usual way.

John Lewis was not entirely happy with all the conclusions in the background paper, but he thinks that the consequences of adopting the suggested strategy would actually be beneficial. Generally, a study is regarded as inadequate if it just achieves the 5% level and has a confidence interval that just excludes zero. To convince doctors, sometimes mega studies are needed where the confidence interval lies miles away from zero and with a really small pvalue. To convince regulators, there is probably an element of that there too, and the two or more pivotal studies question which raises its head from time to time may be because actually $p = 0.05$ is not enough so people want to see two multiplied together and thereby have a smaller

Discussion Papers: A European Concept for Good Statistical Practice

p-value. The nice thing about the suggestion is that it concentrates on doing studies big enough to have narrow confidence intervals and small p-values and therefore, may get researchers back into a world where one pivotal study will be enough.

Robert O'Neill had concern since for many clinical trials in life-threatening diseases *any* difference away from zero is generally felt to be useful. For nonlifethreatening diseases it might be more important to consider the concept of a minimal relevant difference. In interim analysis trials the trial should only be terminated early when there is less information if it is pretty certain that the true difference is far away from zero. Regarding the point Johi~ Lewis was talking about of what it takes to convince the medical community and the regulators, Robert O'Neill responded that for most drugs that are approved, the point estimate of the effect does not drive the extent to which that drug is used by doctors. It may be a medical practice question that may require this estimate to be relatively far away from zero, but in a regulatory context, a point estimate very far away from zero is not likely to happen unless it occurs in a meta analysis context where it is being obtained from three or four studies that have been done.

Wolfgang Kopcke commented from the floor that in an interim analysis that stops early, in most cases the effect is very much overestimated and by group sequential methods

construction of confidence intervals will be complex.

Joachim Röhmel responded to Robert O'Neill's comments that if the confidence interval is totally on the left side of zero and within the equivalence range, that is, at the same time the drugs are equivalent but the test drug is inferior to the control, then in life-threatening diseases the inferiority would be all that matters in the study. He agrees that it might be sensible to define a difference, which must be excluded by the confidence interval approach in a clinical study, but he would

not call it the clinically relevant difference. He pointed out the distinction between superiority and equivalence trials where in both cases CIs are used but two-sided in superiority and one-sided in equivalence, why not use one-sided in both cases so that there would not be any difference any longer?

NN responded from the floor to John Lewis's comment that if a drug lowers highly statistically significantly the blood pressure by 3 mmHg, the company will be able to sell it to doctors. NN also commented that a new drug could very well be equivalent in efficacy by having at least 90% efficacy as compared with a control drug but beneficial because of better safety; here the minimum relevant difference and the difficulty in defining it comes in.

Robert O'Neill expressed concern with the picture presented by NN because if the point estimate is on the negative side one is essentially allowing for the drug to be truly 5-10% worse and one has to live with that. If there is an equivalence trial, say an AIDS trial with a mortality endpoint, the zone of equivalence is very important and it is not the same zone of ~20% in bioequivalence, one can drive a truck through that, but people think that the blood level does not impact the clinical outcome that much. In a clinical outcome situation one is really looking at a very narrow interval and so the suggestion may be to modify this to have an asymmetric range with a limit on the side of being worse of maybe something like 5% because one has to live with that as the truth, it may be up to 5% less effective. Karsten Schmidt responded that this was exactly why he started by emphasizing that limits should not be symmetric so that for life-threatening diseases the lower limit could be zero or at least very close to zero.

Stephen Scnn pointed out that trials do not just have power of, for example, 60% versus 80%; power is a function of the true difference. At the stage where one has to decide how many patients to recruit one

487

iii

*Discus
sion
Coordi
nators*

must commit to some sort of postulated difference, but that does not mean that this particular difference has any relevance in reporting the trial, because when reporting a confidence interval different doctors will inevitably make decisions as to whether they consider this treatment equivalent or not. The problem is that regulatory agencies must make formal decisions as to whether to approve the drug or not. Here the point raised by Joachim Rdhmel is extremely interesting. Regulators seem to have accepted that regulators' risk is 5% in equivalence trials but they require industry to run two-sided tests for superiority trials with a regulators' risk of 2.5% since they will never allow industry to register a drug that was statistically significantly worse than placebo.

John Lewis raised another issue which is bugging him. Patient by treatment interaction effect is a real phenomenon which researchers tend to ignore when applying a fixed treatment effect model. Even though there is a treatment that on average does not exceed some critical value, there may be a subset of patients who were perfectly adequately treated by any definition that might have been chosen.

Robert O'Neill responded that he would like to see on one page some graphical display of confidence intervals broken up by subgroup across all studies to see the consistency and heterogeneity across studies and across subgroups along the lines of meta analysis graphs, and to see how wide those confidence intervals become. His guess would be that there are

going to be very few instances where the CIs do not intercept zero.

Deborah Ashby again pointed out the value of these relatively straightforward summaries of data. She quite often finds herself trying to summarize and sketching out CIs wondering why she should be doing it when the company statisticians are there. Even at the appeal stage, good summaries may be missing and it may be difficult even to figure out how many studies there are. If companies want the regulators to make the right decisions, they should provide good summaries as well as details of individual studies.

Joachim Röhmel reported his experience of having help from his colleagues who only give him the few pivotal studies in a package to look at, this way limiting his work to two or, in rare cases, three studies and perhaps 40 books or so, a maximum until now of 250 books. He has never missed summary graphs but thinks they would be nice to have.

Robert O'Neill responded that he was disappointed that Deborah Ashby was not getting applications that were relatively well laid out in terms of summarized statistics. In the United States, they would probably kick that back and say this is not good enough. This is an issue that could easily be resolved and there are actually procedures in place to allow that to occur. Regarding reporting of clinical studies there is currently an International Conference on Harmonization (ICH) document on that. ICH has been in existence about four years to look at standardization and harmonization of drug development and regulatory procedures in Japan, the European Union, and United States.

Karsten Schmidt summarized the discussion by concluding that apparently there is consensus about statement 4 regarding retrospective power calculations. Most of the reservations concerning statements 1, 2, and 3 largely originate from the concept of minimum relevant difference, which is not well understood and apparently bothers many people. Thus, a great deal of the discussion was on the concept of minimum relevant difference in spite of the fact that it was explicitly mentioned in the introduction that the concept should not be discussed. It has been mentioned that p-values could be informative in some cases, but discussants seem to agree that confidence intervals are a preferable way to present results. Therefore, if this point of view on CIs continues, then sooner or later CIs must be incorporated more directly into the regulatory decisions

488

F

Discussion Papers A European Concept for Good Statistical Practice

and thereby into sample size determination.

**DISCUSSION COORDINATOR'S
RETROSPECTIVE COMMENTS**

This topic has been discussed at length in the literature, and an exhaustive list of references will not be presented here. However, it will be valuable to give some. The topic of having confidence intervals rather than p-values has been discussed extensively (2—12). Regarding the idea of equivalence testing in active control studies, three papers are mentioned (1,13,14), and for the CI or shifted null hypothesis approach in superiority testing four papers are given (15—18).

The main reason why the panel seems not to fully agree to proposed consensus

statement 1 of decision criteria based on confidence intervals seems to be difficulties in accepting the MIREDF concept. It was known in advance that people can have very different opinions of the size in concrete cases, and the decision is often based on implicit and arbitrary judgments, and therefore, it was suggested that the assumption be to agree on the MIREDF range being established one way or the other. What is really surprising is that among statisticians the concept is not well understood. It is hard to understand that there should be any discrepancies between the minimum relevant difference, the smallest clinically meaningful difference, the difference worth detecting, the smallest clinically worthwhile difference, the difference one does not wish to miss, the acceptance range for equivalence, and so forth. These are all synonymous, and nothing will be gained by inventing yet another name as proposed by some discussants. It seems to be equally obvious that the MIREDF has nothing to do with either the difference, the difference one would claim, or the expected difference. The paper by Spiegelhalter and Freedman (17) where this is clearly explained is recommended.

Actually, the FDA (19) requires that the MIREDF be prespecified in the protocol. In Section III.B.7.b. this so-called "delta value" is defined as "a difference between treatments that would be considered clinically meaningful," whereas in Section III.B.9.d. 1 .c "clinically" is not mentioned but only "a meaningful treatment difference."

To be of any value the MIREDF range must be acceptable to the regulators who should, on the other hand, not care much about the applicant's expectation used for dimensioning the study. If, as Sylvain Durnleman suggests, the companies apply overly optimistic expectations, they will soon learn that the criteria for approval will not be met and the drugs will not be approved, which will certainly teach them to do better.

Equivalence must be interpreted with great caution. Equivalence in no way means interchangeability, but refers only to a particular efficacy variable in a particular study design where a treatment difference of negligible size has been demonstrated. In superiority trials some problems with generalizability also exist, but they have not yet prevented such trials from being used extensively.

Willi Maurer's interpretation "that because the power attached to the MIREDF is low, it does not matter to miss it and why not set it to zero and go on the usual way" is confusing. The idea is contrary. By shifting the null hypothesis the power to conclude correctly that the difference is within the MIREDF range is guaranteed to be at least 95%. In other words, the regulator's risk (ie, the probability of letting the company claim superiority when the true difference is of no clinical relevance) is controlled at 5% or less. If the true difference exceeds the upper limit of the MIREDF range, then the company can attach any power it wants by increasing the sample size.

Regarding the use of the CI approach in equivalence trials, Joachim Röhmel is the coauthor of a paper (14) proposing ex

489

L

III

actly what is proposed in the background paper. Some of the words are different, for example, MIREDEF is called acceptance range, but the meaning is the same. Joachim Röhmel now suggests, seemingly supported by Stephen Senn, that “one one-sided intervals” be used or the regulator’s risk be restricted to 5% in superiority trials as in equivalence trials, and this is identical to consensus statement 1 of employing 90% CIs in both cases.

The MIREDEF range will, of course, have to vary, depending on the drug and the parameter considered. It will also have to be prespecified whether one is conducting an equivalence or superiority study. The option of choosing the lower limit at zero, or quite close to zero, must relieve most of the concern expressed in the discussion. Except for the proposal (consensus statement 1) of employing 90% CIs not only for equivalence trials, but also for

- superiority trials, the proposal is actually so flexible that usual nonshifted null hypothesis superiority testing is the subset where the MIREDEF range collapses to a single point equal to zero (or one if, eg, odds ratios are considered) and thereby covered by the proposed approach as well. Therefore, no one is forced to change strategy as long as a justification is given according to consensus statement 3.

Gary Koch, Uwe Ferner, and Wolfgang Köpcke seem to like p-values because CIs are more complex to calculate and require choice of an appropriate parameter, metric, scale, or whatever it is called. In a randomized clinical trial, a small p-value only indicates that it is unlikely that the result under consideration could have been produced by randomization only. This is useful, although often not sufficient, information. Also, p-values require a choice of metric since to be calculated some test statistic or “discrepancy measure” that quantifies the extent to which the results deviate from the null hypothesis must be chosen.

The problem with p-values is that to interpret them properly they must be combined with some further information on

Discussion Coordinators

the study. In mega size studies, small p-values can occur even when the treatment effect is of a negligible magnitude that is completely clinically irrelevant. Therefore, small p-values do not necessarily mean that the CI lies miles away from zero. Actually, what an experienced statistician does when looking at p-values is combine them with information on sample size, null hypothesis, test statistic, and so forth to form in his mind something that is pretty much like a confidence interval to be able to interpret the p-values in a reasonable way. So why not aim directly at CIs even when it is more complex to do so? In most cases, a CI can be determined at least by Monte Carlo simulation and trial and error methods trying to find which alternative null hypotheses can be accepted given the observed value of the test statistic.

It is very satisfying that the discussants have all agreed on consensus statement 4, and hopefully, this will have an impact on future revisions of the FDA guideline (19), in which retrospective power calculation is suggested as an option in Section III.B.9.c.2.g and as a recommendation in Section III.B.9.d.1.c.

The issues discussed here are actually elementary and fundamental for the statistical analysis of every clinical trial forming a basis for drug approval. The statistical profession should be ashamed that this has not been sorted out a long time ago. It ought to be spelled out in every guideline and elementary textbook on statistics in clinical trials. Confusion about

the meaning of words is not a good excuse.

REFERENCES

1. Temple R. Government Viewpoint of Clinical Trials. *Drug Inf J*. 1982;16:10—17.
2. Simon R. Confidence Intervals for Reporting Results of Clinical Trials. *Ann Mt Med*. 1986; 105(3):429—435.
3. Rothman KJ. Significant Questing. *Ann Mt Med*. 1986;105:445—447.
4. Langman MJS. Towards Estimation and Confidence Intervals. *Br Med J*. 1986;292:716.

V

I.

Discussion Papers: A European Concept for Good Statistical Practice

- 5.7 Gardner Mi, Altman DG. Confidence Intervals Rather than p-Values: Estimation Rather Than Hypothesis Testing. *Br Med J (Clin Res)*. 1986; 292:746—750.
6. Berry O. Statistical Significance and Confidence Intervals. *Med JAustralia*. 1986;144:618—619.
7. Fleiss iL. Confidence Intervals vs. Significance Tests: Quantitative Interpretation. *Am J Pub Health*. 1986;76(5):587.
8. Bulpitt Ci. Confidence Intervals. *Lancet*. 1987; 1:494—497.
9. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts submitted to Biomedical Journals. *Ann mt Med*. 1988; 108:258—265.
10. Bailar JC, Mosteller F. Guidelines for Statistical Reporting in Articles for Medical Journals: Amplifications and Explanations. *Ann mt Med*. 1988; 108:266—273.
11. Braitman LE. Confidence Intervals Extract Clinically Useful Information from Data. *Ann mt Med*. 1988;108:296-298.
12. Gardner Mi, Altman DO. Statistics With Confidence—Confidence Intervals and Statistical Guidelines. In: *British Medical Journal*. London, United Kingdom; 1989.
13. Dunnett C, Gent M. Significance Testing to Establish Equivalence Between Treatments, With Special Reference to Data in the Form of 2 x 2 Tables. *Biometrics*. 1977;33:593—602.
14. Garbe E, Rohmel 1, Gundert-Remy U. Clinical and Statistical Issues in Therapeutic Equivalence Trials. *Eur J Clin Pharmacol*. 1993;45:1—7.
15. Schwartz D, Flamant R, Lellouch 1. *Clinical Trials*. London, United Kingdom: Academic Press; 1980 (1970 in French).
16. Freedman LS, Lowe D, Macaskill P. Stopping Rules for Clinical Trials Incorporating Clinical Opinion. *Biometrics*. 1984;40:575—586.
17. Spiegelhalter DJ, Freedman LS. A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion. *Stat Med*. 1986;5:1—13.
18. Victor N. On Clinically Relevant Difference and Shifted Null Hypotheses. *Meth Inf Med*. 1987; 26:109—116.
19. FDA. *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*. Rockville, MD: U.S. Department of Health and Human Services, Food and Drug Administration; July 1988.

5

PLACEBO-CONTROLLED TRIALS AND ALTERNATIVES

DISCUSSION COORDINATOR: MIKE COLLINS
Hoechst Roussel Ltd.

BACKGROUND

Placebo Controlled Trials

1. ONE OF THE key decisions in the design of a clinical trial is the choice of “control” treatment. Efficacy for a new drug may be established by demonstrating superiority against placebo. Randomized, double-blind, placebo-controlled trials are the “gold standard” for the demonstration of efficacy of a new drug, as failure to demonstrate a clinically relevant statistical difference against placebo (assuming appropriate statistical power) may be due to noneffectiveness of the drug under investigation. It is often desirable, however, to use an accepted standard therapy as another treatment arm within a placebo-controlled trial. This standard treatment arm will usually only be used for the demonstration of the validity of the trial itself and not in direct comparison with the new drug. The choice of comparator may also be related to the objective of the trial, for example, an explanatory trial would normally include a placebo control group while a pragmatic trial would conventionally use a standard comparator, if available. There may be situations where there is no accepted standard treatment available or in a multicenter clinical trial there may be no clinical consensus regarding the choice of a standard treatment. It is often difficult to determine the magnitude of a clinically relevant difference.

There are often major ethical concerns, however, in giving placebo treatment to patients in clinical trials. The magnitude of these concerns varies both within and between individual countries in Europe. It often depends upon the disease under investigation (eg, chronic/acute, life-threatening), the availability of a therapeutic alternative, the route of administration, and the duration of treatment. The number of patients receiving placebo may be minimized by using unequal randomization (eg, drug : placebo: 2:1 or 3:1, this will result in a loss of statistical power) or by the specification of early stopping rules to demonstrate a clear clinical benefit as early as possible. Do placebo-controlled trials reflect clinical practice?

Another concern in the design of a placebo-controlled trial is maintaining blindness, especially for the unbiased reporting of adverse events (AEs) and subjective efficacy rating scales (eg, clinical global impression scales). This is particularly important in therapeutic areas where high placebo response rates are seen. Compliance to placebo treatment is often difficult to document and can be confounded with high dropout rates due to lack of efficacy; conversely, dropout rates under active treatment may be related to tolerability as well as lack of efficacy.

What are the Alternatives?

1. *Demonstrating efficacy by showing superiority against an accepted standard treatment* – Is placebo required as an internal standard? An accepted standard must be chosen and validated.

2. *Demonstrating efficacy by showing therapeutic equivalence against an accepted standard treatment*—Although statistical methods are widely available for equivalence testing, problems still remain, for example, choice of standard, “noneffectiveness” of the standard, and changing clinical practice giving new standards. Various other factors such as insensitive variables, high variability, improper conduct of the trial, expectation of equivalence, and failure of the trial itself may minimize treatment differences. Furthermore, not only equivalence itself must be demonstrated but also some clinically relevant change from baseline. One needs assurance that the active control is performing better than placebo from information external to the trial. Therapeutic areas where high placebo response rates are known may require an internal placebo group for validity!
3. *Demonstrating efficacy by showing a clear dose response relationship* – In Ethics this case, the lowest dose of a new drug will be a quasi-placebo. This does not differ principally from the case of a real placebo. Superiority of the therapeutic dose versus the lowest dose must be shown. Choice of the doses and the number of doses to be assessed may be problematic.
4. *Explaining efficacy by unquestionable pharmacological effects*—Is this really acceptable?

DISCUSSION

The presentation opened up a number of issues during the discussion which are summarized below.

Placebo Response There are very few trials which actually (according to the standard of the randomized clinical trial) prove that there *is* a placebo response. Usually what people mean

Discussion Coordinators
by placebo response is some difference from baseline in the placebo group. But this, in fact, is no proof that there is a placebo response as it may be due to regression to the mean, trend effects, or observer bias, which do not imply a placebo response by the patient. Placebos are used to try to control for these types of effects. It may be highly appropriate to use a placebo when a high placebo response is expected. There is no such thing as a baseline condition for a patient. Patients move through time during which they do, in fact, receive all sorts of concomitant “treatments” – regular intake of food and perhaps concomitant medications as well. Therefore, it is not true that patients must be deprived of all treatments in order to be able to assess the effect of a new experimental treatment. The increased precision and sensitivity of measurements (eg, 24-hour blood pressure recording) should minimize placebo effects. Placebo effects are less often seen in long-term studies.

A critical issue is the suggestion that all patients should be entered into randomized trials in order to find out more about therapies all the time. There are some ethical arguments for this saying that future patients benefit from randomized trials, therefore, they should also play their part in contributing to the stock of future knowledge. But now, suppose that what is thought of as the ideal randomized trial is a placebo-controlled trial in which the patients receive either the new experimental drug or a placebo. And suppose that it is thought that all future patients should be entered into randomized trials. What this means is that as soon as a drug is developed, it will never be sold, because all future patients will either be receiving a new experimental treatment or they will be receiving a placebo! But certainly, on empirical grounds there are many instances when a placebo is not an unethical option for the patient. There are examples where

- it is actually better to be on the placebo for a life-threatening disease (eg, recent congestive heart failure trials). Often, the best thing is to just be in a clinical trial. One of the ethical difficulties faced with a placebo is also caused by the fact that this issue of placebo-controlled trials is primarily discussed within scientific circles, between statisticians and clinicians, and that it is not well-exposed to the public and actually to the people who are going to receive placebo. The profession should make every effort to make sure that the general public is part of the discussion because it is not only a scientific issue.

Pharmacological Effect

Showing a pharmacological effect as an alternative type of study is rather a vague concept. The definition of a pharmacological effect must be clear (eg, bacteriological response). Are these surrogate endpoints? It will be important to demonstrate and justify that pharmacological effects are directly related to some more important endpoint from the patients' point of view. The balance between risk and benefit can also help to clarify this issue.

Dose Response

If a low dose of a drug is used, effectively as a placebo, then either it is a placebo (in which case nothing is learned about the dose response) or it is not (in which case one is not sure how active it is relative to placebo). So the only way to solve that is by having two doses and a placebo. A multiple-dose study cannot stand alone without some placebo-controlled studies elsewhere. Another aspect of the dose response design is where the doses lie on the dose response curve; a series of studies must be planned to describe the curve. Different individuals have different dose-response curves. So one concern with active-controlled trials is that one never knows whether one is testing one treatment against the other treatment's optimal dose.

Unblinding

The risk of unblinding a trial because of the safety profile is not only encountered in placebo-controlled trials but also in active-controlled trials where the safety profiles of drugs are different.

Choice of Comparator

It may be difficult to choose an active comparator when one wants to run an active-controlled study, for example, in a multinational setting, because there is no unique accepted standard. A pragmatic approach seen particularly in the pharmaco-economic setting is the so-called naturalistic design where one well-identified treatment is compared with the best care available in each country or center.

Other Designs

The crossover design at least gives each patient both active and placebo treatment but may be

problematic in certain indications. Another design is the rescue design where everyone gets the drug and one randomly withdraws the drug from half of the individuals, and then defines an endpoint which essentially requires rescue medication (eg, pain studies). Another design which is potentially problematic is the responder (enrichment) design where essentially those individuals who respond or have either a high or a low placebo rate are screened out and then responders are randomized. Adaptive designs may also be useful. A further way in which clinical trials could be run especially for lifethreatening diseases and where an effective treatment is already available is in the form of "add-on" therapy. One does not hold the baseline effective therapy but one offers in addition half of the patients

a
495

496

placebo and the other half an experimental drug.

Terminology

The term placebo should be reserved for randomized trials and, in fact, only for randomized sections of the trial. With a randomized trial, one can show the patient the protocol and every detail of the trial and say: "What you will receive, if you agree, is a placebo or an active treatment. And this is the probability of your receiving placebo." On the other hand, people

Discussion Coordinators

also use the word placebo for run-in phases of trials. If one wants the patient to be blind as to what he/she is receiving in the run-in phase, then one cannot, in fact, share with him/her every detail of the protocol. The same argument applies to any trial in which placebo always appears in a particular sequence (eg, washouts). It may be useful to use a straightforward term such as "inactive" rather than placebo elsewhere. It is also important to distinguish a Hawthorn effect and a placebo effect; namely, the effect of being in a clinical trial rather than specifically of being on a placebo.

Drug Information Journal, Vol. 29, pp. 497-502, 1995 Printed in the USA. All rights reserved.

0092-8615/95

Copyright © 1995 Drug Information Association Inc.

STANDARD OPERATING PROCEDURES

DISCUSSION COORDINATOR: ANNETTE ROBERTSON
Zeneca Pharmaceuticals

WORKING GROUP MEMBERS:

ANNE WILES
Brookwood Statistics Ltd.

JOHN ALEXANDER

Syntex Research

PETER CLARK
Wellcome Research

MICK GODLEY
Zeneca Pharmaceuticals

DENNIS LENDREM
Sterling Winthrop Research

DISCUSSION

GOOD STATISTICAL PRACTICE is continually developing. Although the overall objectives remain the same, the methods by which those objectives are achieved are constantly changing in line with technical developments, changing business or regulatory needs, and technological advances. It follows, therefore, that statistical standard operating procedures (SOPs) must be “live” documents if they are to actively encourage good statistical practice. Consensus on the following issues is key to achieving a common approach to statistical SOPs.

1. *What is the purpose of SOPs?* – Is it compliance with GCP and other regulatory guidelines or business driven with the emphasis on regulating internal quality by setting operational standards (eg, in software development) since rework is costly?

It is both. Although the driving factor in developing SOPs initially has often been to meet regulatory requirements, this emphasis is changing and many organizations now see a clear business need for SOPs. They should be an aid to process improvement.

2. *Should an SOP be a statement of principle or a detailed set of instructions?*— A detailed set of instructions is helpful in the training of new staff and would probably help the auditor. It is, however, inflexible, difficult to maintain, and inhibits professional freedom.

It is suggested that SOPs should be statements of principle with appropriate options and topics for consideration indicated. It may be appropriate to support these with more detailed guidelines on working practices in some instances such as to assist with staff training.

3. *Interfaces and Harmonization*

a) It is reasonable to suggest that:

- The statistical contribution is part of a multidisciplinary process and consistency of operation and expectation is desirable, and

497

498

- A statistician should review any presentation or interpretations of the data analyzed for statistical validity.

If the above are to be achieved the mechanism by which statistical SOPs interface with SOPs from scientific, medical, data management, marketing, and registration groups within the same organization needs careful consideration. It is suggested that collaboration is essential to the quality of the overall process.

- b) Should SOPs be harmonized at the detailed level or at the level of “statements of principle”:

- In international pharmaceutical companies?
- Between a sponsor company and a contract research organization?

It is suggested that the pragmatic solution is to work to common overall quality standards and principles but that achieving common working practices may be counter-productive. Changing working practices purely to achieve harmonization of SOPs may produce a reduction in quality rather than an improvement, as it may involve considerable culture changes in some parts of the organization. Some areas such as global database requirements and regulatory requirements, however, must be considered for harmonization.

4. *Compliance with SOPs—How can it be encouraged?*

a) There must be a single standard that meets both business and regulatory needs.

- b) SOPs must be accessible with clear and useful content.
- c) The documentation system must be well managed (eg, good version control) – should the documentation system be electronic or paper?
- d) Culture/People
 - SOPs should be introduced positively into an organization as part of a quality system,
 - Staff must be trained in SOPs and

Discussion Coordinators

- necessary technical areas regularly,
- Staff should be involved in the production/writing of SOPs to help gain their commitment,
- Staff must be given regular feedback from quality control (QC) checking procedures, and
- Formal mechanisms must exist for authorizing and handling deviation from SOPs.

The above requires considerable commitment of resources to ensure regular review and update of the content of the SOPs, management of the documentation system, training, and so forth. These factors all encourage compliance with SOPs and are concerned with the *prevention* of problems. But where has the *audit* gone?

QUALITY CONTROL OF STATISTICAL CONTENTS OF REPORTS

1. *Quality Control and Quality Assurance: Definitions*—The quality assurance system as defined in European community good clinical practice includes quality control as one element. QC is defined as those operational activities designed to ensure that a company's products (tables, listings, graphics, analyses, interpretation, reports, datasets ...) meet the requirements for quality.

The objective of a quality assurance (QA) audit is to ensure that QC systems are functioning correctly. Thus, statistics units should not depend on the audit work of QA groups to fulfill a QC role; errors detected by retrospective QA audits mean costly rework and delay. Statisticians should ensure they have adequate QC systems to build quality into the product.

2. *Specification and Management of QC Procedures and Tasks*— The specification of analyses and reports must be

e

Discussion Papers: A European Concept for Good Statistical Practice

agreed upon between “customer” and “supplier.” Frequently, this will be achieved by reference to internal reporting standards which should be documented and accessible.

QC activities must focus on means of preventing nonconformance to the specification (by, eg, cross-checks, test data) as well as detecting errors through inspection. QC procedures in routine use must be specified in SOPs. The objectives of and methods to be used for each QC task must be stated; “review the report” is not an adequate description. Check lists should be used.

The skills required for each QC task must be specified and training provided. Statistics management should ensure that it is resourced to meet QC needs, and not regard them as activities that can be squeezed in around “normal” work.

3. *Some Key QC Activities*

- a) Protocol and CRF Review: peer review by experienced statistics staff, with a check list to confirm that all aspects of the design and proposed analysis methods are appropriate to the study objectives and the type of data.
- b) Software: it is not necessary to validate software products (eg, SAS) except when updated product versions are introduced or features are used for the first time. Programs written in SAS or other packages should be validated by the documented use of test data. Standard macros should be maintained in a change control environment.
- c) Analysis plan/dummy analyses and reports: listings, tables, and graphics should be checked for accuracy and conformance. Proposed analyses, tables, and figures must be confirmed by expert peer review. A “customer” review of the material should be undertaken.
- d) Final report components: data listings should be checked for accuracy prior to analysis, to avoid rework. Tables, graphics, and analyses must be checked for completeness and accuracy (data, titles, footnotes, annotations); they must match tables of contents and should reflect planned analyses. An expert reviewer should check patient evaluability assessments and confirm that test results and estimators are presented and interpreted correctly.
- e) Statistical and study reports: any stand-alone statistical report must be independently checked for readability and correct language use as well as typographical accuracy. If the study report is produced separately (eg, by a medical writer) the study statistician must check to ensure correct interpretation of results. In addition, there must always be independent expert statistical review to ensure that final report conclusions are consistent with the results of the analyses.

The discussions following the presentation of the key points in the paper were mostly based around SOP production, contents, and usage, including aspects of quality assurance and responsibility. SOPs are becoming an essential part of the pharmaceutical business, and production of company SOPs has been a key activity in recent years. On the regulatory side, the existence of SOPs is rare, and this was felt by the panelists to be an area which would benefit from having SOPs in place. Some units in the FDA have SOPs, but not all, although staff manual guides have been available to FDA reviewers for 15 years. It was anticipated that the FDA will produce documents detailing the review process, and that an overview SOP was required to increase consistency between the reviewing divisions. The point was made that if regulatory authorities are going to use external expert reviewers, SOPs were going to play an important part in assuring consistency of review.

499

500

Discussion Coordinators

SOPs are difficult documents to produce, requiring commitment and effort. In order to be effective, they need to be practical, but not too detailed. SOPs usually need to be developed and agreed upon internationally, which can involve a complete review and revision of the current processes. The ownership of these documents needs to lie with upper management to ensure the necessary commitment. When developing international procedures, it was recognized that there may be differences in the way in which these were adhered to. A certain amount of compromise in the local processes involved may be required, but the principles outlined in the SOPs need to be achieved and maintained.

It was generally agreed that SOPs should be about the principles of what should be done and when, as well as outlining the responsibilities. SOPs should not include the detail of how to carry out tasks —these more detailed technical processes could be covered by additional documents to be used as guides, particularly when the SOP is international with the processes beneath it differing between countries. The inclusion of check lists within an SOP, however, was seen as helpful, and necessary.

One of the key elements of an SOP is to clearly define the interfaces between different working groups and departments within and outside the organization. The SOP can be used as a contract between the respective parties, and should be developed and agreed upon by those groups to whom the SOP will apply.

The amount of detail to be included in an SOP was a subject of much debate, and brought in the question of an audit against documents. A document which outlines the required processes in general terms, without specific details on responsibilities, timings, and requirements, would be unlikely to produce problems if used as an audit tool by a regulatory authority. This SOP, however, would not be a useful document to the company, and would not provide consistency. At the other extreme, too much detail in an SOP results in that procedure being much more difficult to comply with, and may cause problems under audit.

It was noted that the draft CPMP guidelines include reference to validated computer software. One opinion was that more detail needs to be included on this topic, as mistakes are made during programming, particularly if a program is produced to perform a new statistical procedure. In this situation, validation is important. It was also noted that apparently validated software may not produce the right answer, and that major analyses should be compared using different software.

A subject which received major attention was the sign-off on protocols, study reports, and submissions by statisticians. Some members of the panel saw this as a vital part of the statistician's responsibilities, particularly given that statisticians are usually the most objective people taking part in the process of drug development, and the only members trained in the scientific interpretation of data. There was, however, a view that there are insufficient experienced statisticians with a perspective of drug development and the production of an integrated summary of efficacy and an integrated summary of safety to be able to make this a requirement of regulatory authorities. This area was seen as a possible niche for contract organizations.

There was concern that should the sign-off of clinical reports and higher level documents be made a requirement in a regulatory guideline, pressure may be applied by companies on statisticians to sign. What would the consequences be should they refuse to do so? It was felt that the philosophy of the companies needed to change, recognizing the value of statisticians, rather than having sign-off of documents imposed. On a similar line of discussion, it was noted that the introduction of the FDA guidelines had a major effect on the employment of statisticians in pharmaceutical

Discussion Papers: A European Concept for Good Statistical Practice

- tical companies, and on the responsibilities held by these statisticians. There was also comment that in Europe the lack of inspections by regulatory authorities has meant less emphasis being placed on the importance of statistics.

A regulatory member of the panel commented that the statistical review of expert reports would be valuable, as these can contain conclusions which are not in accordance with the findings of the clinical studies. Statistical review could prevent this, and lead to objective data driven conclusions.

SOPs can be used by biometrics groups to train staff on the processes involved in drug development and individual responsibilities. Regular training events were seen as important, and it was noted that dedicated resources are needed to enable this to happen.

Quality control and quality assurance were two areas where the use of SOPs were seen to be vital. Both elements should be built into the SOPs, and both the company and regulatory authorities can audit against the SOPs. SOPs themselves provide documentation of the quality standard expected, and the use of check lists and other forms as part of compliance with the SOPs can provide documentation of quality. There was a suggestion that this may speed up the review process.

A question asked by a member of the audience centered around the adequate training of statisticians to be able to carry the responsibility for a quality product, and compliance with SOPs. This resulted in a discussion which moved away somewhat from the subject of SOPs, and returned to the training of statisticians and the sign-off on clinical reports and packages.

It was suggested that the universities could do more to educate students in the skills needed for specific jobs, and be aware of the areas for which their courses are applicable. It was recognized, however, that it was not possible to do all the necessary training within the university, and that employers have a responsibility, too. A very relevant comment made was that education

never stops, and that self-development should be encouraged. In addition, the university level may not be the appropriate place for this training, as many students do not have a defined career path at this stage. It was suggested that companies should explicitly give responsibility for statistical training to one statistician as that person's main job role. This was seen as the best way of guaranteeing ongoing training. The concept of mentoring was also seen as extremely useful in the work environment, and could provide greater on-the-job training. Placing the responsibility of training staff in the hands of the employer also gives individuals opportunities and scope for development within the company.

The discussion had to be closed due to time constraints, and had the potential to continue for some time. The quality control of statistical contributions was not discussed in as much detail as SOPs, and the discussion became sidetracked somewhat by the issue of training. The following points formed the closing summary:

- Responsibilities need to be clearly defined, and should be addressed explicitly in SOPs,
- SOPs should be produced as the level of principle, and not contain the detail,
- SOPs are appropriate for regulatory agencies, as well as sponsors,
- Check lists were considered to be very useful in terms of clarifying the processes, and as a training aid,
- The fact that software is validated does not preclude quality control of the output,
- There is a need for guidance for the experts involved in the production of expert reports, and in the review of submissions,
- Job-specific training needs to be a company responsibility, and statisticians need to remember that education is an ongoing activity, and

.501

H

k

*Discussion
Coordinators*

°

There was a move toward statisticians being responsible for signing off on major study reports and higher level documents, including the expert report.

In conclusion, the possibility for the pharmaceutical statistician to play a much greater role in drug development was highlighted by the discussion. Efficiency and effectiveness are going to be the key words in the future, and with the objective analytical training that the statistician receives, he/she is ideally placed to take the initiative and contribute to drug development in much broader terms than as just "data analysts."

502

FORMAT/CONTENT OF OVERVIEWS (INTEGRATED EFFICACY AND SAFETY REPORT)

DISCUSSION COORDINATOR: REMY VON FRENCKELL

EFSPi, Bristol-Myers Squibb, PRL, Belgium

WORKING GROUP MEMBERS:

JIM COOPER

Rhone-Poulenc Rarer, United States

HANS ESSERS

Solvay Duphar, The Netherlands

AD THEEUWES

Organon, The Netherlands

INTRODUCTION

The New Drug Application (NDA) Dossier

IN ITS FOURTH part (technical sections), the NDA dossier (Tables 1 and 2) contains a section called "Clinical data" of which the second part is "Integrated summaries" which mainly deals with the efficacy and safety overviews. The integrated summary of effectiveness is a review of the efficacy results showing that they represent adequate and well-controlled studies demonstrating the claimed effect. The integrated summary of safety information is an overall analysis examining all studies together.

The EC Dossier

There are no global summaries specified in the EC dossier (Tables 3, 4, and 5). However, global overviews of efficacy and safety are generally incorporated into the clinical expert report which is to be "a critical evaluation of the quality of the product" and has "to summarize all important data in an appendix."

FROM A STATISTICAL POINT OF VIEW

i~.

2.

503

The FDA “Guidelines for the Format and Content of the Clinical and Statistical Sections of an Application” details the contents of:

1. *The Integrated Summary of Effectiveness Data* of which key points are:

- An examination of study-to-study differences in results, effects in subsets of the treated population,
- Dose-response information from all sources,
- Any available comparisons with alternative drugs,
- Ordinarily, studies with similar controls should be discussed together,
- It is generally not helpful to pool results from individual studies not designed for analysis in that fashion, and

Evidence of long-term effectiveness, tolerance, and withdrawal effects, and
The Integrated Summary of Safety Information of which key information is:

504

Discussion Coordinators

- Overall extent of exposure,
- Grouping of studies (all controlled trials, foreign trials, domestic trials),
- Grouping of events,
- Analysis of adverse effect dose-response information,
- Long-term adverse effects (six months or more),
- Withdrawal effects,
- *Update of safety information on a regular basis.*

The CPMP Note for Guidance “Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products” has a specific chapter (14) on summary or meta-analysis of the results of several trials.

Among other considerations, the section which refers to efficacy states that:

an individual study may not have sufficient power to yield convincing results, and overall summaries of the results of more than one trial may alleviate this problem.

The section referring to safety states that:

The combination of results from several studies will often be appropriate. However, the open-ended search in large databases is fraught with difficulties arising from ‘data-dredging.’ Since the incidence of adverse events is often related to duration of exposure and/or follow-up, use of survival analysis methods will often be appropriate.

TABLE 1

The NDA Dossier: Global Structure

TABLE 2

The NDA Clinical Part: Structure

1. Description/analysis of clinical trials

2. integrated summaries
3. Abuse and overdosage
4. Informed consent and IRB

TABLE 3

The EC Dossier: Global Structure

- Part I. Summary of the dossier and special particulars
- Part II. Chemical, pharmaceutical, and biological documentation
- Part III. Pharmacotoxicological documentation
- Part IV. Clinical documentation

TABLE 4

The EC Dossier: Structure of Part I/IV

- Part I.
 - I A: Administrative data
 - I B: Summary of product characteristics
 - I C: Expert reports Part IV.
 - IV A: Clinical pharmacology
 - IV B: Clinical experience
 - Clinical trials
 - Postmarketing experience
 - Published and unpublished experience
- 1. Application form
- 2. Index
- 3. Summary
- 4. Technical sections
 - CMC
 - Nonclinical pharmacology and toxicology
 - Human pharmacokinetics and bioavailability
 - Microbiology
 - Clinical data
 - Statistics
- 5. Samples and labeling
- 6. CRPs and tabulations

TABLE 5

The EC Dossier:

Structure of the Expert Report

1. Problem statement
2. Clinical pharmacology
3. Clinical trials efficacy
safety
4. Postmarketing experience
5. Other information
6. Conclusions
7. Reference test
8. Information of the clinical expert

Discussion Papers: A European Concept for Good Statistical Practice

TOPICS FOR DISCUSSION

Statistical Topics and General Topics

- Should the same statistical approach be used for both summaries (efficacy and safety)?

- Is a common approach suitable across all indications?
- Should the same statistical approach be applied **in** combining studies as for combining centers in multicenter trials (along the lines of the EC guideline)?
- Are guidelines needed for meta-analyses?
- The exercise of conducting a metaanalysis might itself be repeated several times as the body of trial reports grows. Therefore, the problem of repeated significance testing needs to be addressed,
- Deliberate design of a clinical research program as a series of studies intended for an eventual meta-analysis and interim meta-analyses might be used to determine when the program should be stopped,
- For European regulatory authorities, the reviewers may not routinely have access to statisticians within the health authority,
- What studies should be included? Should findings be adjusted for reporting bias (ongoing studies, low enrollment studies, ...)
- Should there be specific summaries on quality of life and cost-effectiveness data?

505

Drug Information Journal, Vol. 29, pp. 507-508, 1995 printed in the USA. All rights reserved.
0092-8615/95

Copyright © 1995 Drug Information Association Inc.

INTERIM ANALYSES

DISCUSSION COORDINATOR: ANNICK LEROY

Bristol-Myers Squibb, Belgium

WORKING GROUP MEMBERS

M. BUYSE

International Institute for Drug Development, Belgium

M. HARRISON

Upjohn Laboratories-Europe, Belgium

J. M. ZAYAS

Laboratories Almirall, S.A., Spain

ISSUES TO BE considered/discussed related to interim analyses were:

- Reasons for interim analyses: therapeutic effect (none, or extreme), safety or efficacy, "administrative looks,"
- Will the trial be changed in any way based on the results?
- *Could* the trial be changed in any way based on the results? Are assurances from the trial sponsors that no changes will take place based on the results sufficient/acceptable to satisfy ethical/scientific/regulatory concerns, in cases where the trial could theoretically be stopped based on the results?
- Stopping rules: adjustment of significance levels, specification of boundaries, protection of

- overall Type I error,
- Prespecification in the protocol: number of analyses, timing of analyses, stopping rules, and variables affected by the interim analysis,
 - Blinding (who should be blinded: statistician, company, investigators?, type of blinding: complete blinding, A vs. B presentations; blinding of safety vs. efficacy),
 - Effect of *unplanned* interim analyses on trial conduct
 - Resizing of the study on the basis of interim calculations of, for instance, variance estimates, Usefulness, role, and composition of data monitoring boards: is the appreciation different if an interim analysis is performed by the sponsor or an external independent organization,” and does this depend on the type of interim analysis?

ADMINISTRATIVE LOOKS

Administrative looks should not normally affect the conduct or method of analysis of the trial. The objective of such looks would include a check of recruitment rate, database set-up and cleaning, and protocol compliance. The draft European guidelines state, however, that there can be no *possibility* of changing the trial. What happens if centers are dropped due to slow enrollment, for example? Is this to be considered a change in the design of the trial? What measures are permitted to increase the accrual rate?

SUBSTANTIVE LOOKS

Substantive looks have a possibility of changing the conduct and/or methods of analysis of a clinical trial based on the results.

.507

*Discussion
Coordinator
s*

508

- The traditional purpose of an interim analysis: stopping the trial due to either substantial treatment effect (either efficacy or safety), or negligible treatment effect (efficacy),
- Without adjustment of p-values, repeated testing at a significance level of α leads to an increased probability of rejecting H_0 . How the adjustment of the p-values is done is crucial and, especially if interim analyses are foreseen, should be discussed in complete detail in the protocol. Some of these are briefly mentioned below,
- Reasons for stopping a trial: serious adverse effects; larger than expected beneficial effects of one treatment; and negligible possibility of a statistically significant difference between treatments, even if the trial were continued to its normal term (stochastic curtailment),
- Another possible outcome of a substantive look at the data may be to extend a trial,
- Stopping rules are basically concerned with “how much” of α is “used up” per look at the data. Which method is preferable in practice? The α spending function of Lan and DeMets is increasingly popular due to its flexibility and generality. Two of the more common methods of doing this are the Pocock method and the O’Brien-Fleming method (which are special cases of the Lan and DeMets methodology). The main difference between the two is that Pocock uses the same adjusted α (say, α') for each look, while O’Brien-Fleming uses an increasing α' per look. Overall, if there are large treatment differences, Pocock is better; with small treatment differences, O’Brien-Fleming is better. A related issue to this is that the sample sizes will have to be increased somewhat to maintain power when there are multiple looks at the data,
- Is availability of software a limiting factor?

- What is the acceptability of these approaches by the regulatory authorities?
- In some circumstances, sequential designs can be preferred to group sequential designs.

Drug Information Journal, Vol. 29, pp. 509-510, 1995 printed in the USA. All rights reserved.
0092-8615/95

Copyright © 1995 Drug Information Association Inc.

THE ISSUE OF MULTICENTER TRIALS: CAN ONE MULTICENTER TRIAL BE AN ALTERNATIVE TO THE CONCEPT OF “AT LEAST TWO PIVOTAL STUDIES”?

DISCUSSION COORDINATORS

ALEC VARDY

Innovex, United Kingdom

STEFAN DRIESSEN

N.Y. Organon, The Netherlands

MARCO GIRELLI

Glaxo SpA, Italy

JOHN LEWIS

Medicines Control Agency, United Kingdom

FDA “GUIDELINES FOR the Format and Content of the Clinical and Statistical Sections of New Drug Applications” (I) state that the approval of a new drug requires substantial evidence of effectiveness. Quoting from these guidelines, “(this) requirement ... has been interpreted to mean that the effectiveness of a drug should be supported by more than one well-controlled trial and carried out by independent investigators. This interpretation is consistent with the general scientific demand for replicability.”

From within the pharmaceutical industry, many argue that the demand for replicability can be satisfied at least as well (if not better) by a single, generally multinational, multicenter study, provided, of course, that the overall result is positive and that there is little evidence of a clinically meaningful interaction between the treatments applied and the centers applying them. In addition to replicability over investigators, such a trial might also include other sources of heterogeneity, such as different countries, different types of patients, different disease severities, and so forth. Is it seen as desirable that pivotal trials demonstrate a clear treatment effect against a background of realistic heterogeneity? Can such a multicenter trial be an alternative to the concept of at least two pivotal studies in the eyes of the FDA and, if so, under what conditions? Further, how close are other regulatory bodies to “harmonization” with FDA on this issue?

The FDA guideline goes on to admit that there have been cases where a single particularly persuasive study has been accepted, but confirms that these were “exceptional

circumstance? (eg, a study considered unrepeatable on ethical grounds). Troendle (2) has suggested that FDA guidelines would be met if it was specified in the statistical analysis section of the protocol that the centers from a multicenter trial would be divided into two mutually exclusive sets of centers. Nevius (3) has described a four-point proposal for assessing statistical evidence in a single multicenter trial after the trial results are available.

While these suggestions seem at first sight to represent a softening of the requirement written into FDA guidelines, in practice they both depend on the multicen

509

510

Discussion Coordinators

ter study being overpowered, or on sheer good luck. In implementing Troendle's idea, a sponsor might take the precaution of increasing the overall study size in order to reduce the chance that the more positive centers might fall into one set and the less positive into the other, resulting in this second set failing to achieve the desired critical level of significance. Nevius' proposal is designed to be used where a study is found to be overpowered owing to better than expected efficacy, smaller variance in the response, or better recruitment.

Even if neither idea represents an approach of practical value to the pharmaceutical industry, both raise the question of exactly what is meant by the scientific demand for replicability, and both appear to place a liberal interpretation on the concept of "independent investigators."

The pharmaceutical industry actually applies Troendle's philosophy, but generally under another guise. A group of investigators, on occasion from the same area of the same country, is split into two subgroups. The subgroups then participate in a protocol identical except for its number, in an attempt to create two "independent" trials, and hence, claim replicability if both trials were to give positive results.

Contrast this with a single, "adequately-powered," multicenter trial involving several investigators from different countries, conducted to a high degree of excellence, showing a positive result overall which is highly consistent across centers.

What makes the former more acceptable than the latter? In what way is the former better evidence of replicability? Where is the greater degree of independence?

The multinational, multicenter trial, despite using the same protocol, is likely to include a greater degree of realistic heterogeneity. Clearly, some agreement would need to be reached about when the conclusion of consistency across centers is valid (qualitative versus quantitative interaction, the importance of single "rogue" centers, etc.). Clearly, also, the question will need to be raised as to whether a trial with many small centers will be acceptable, or whether credence would only be given to a trial with a small number of large centers. However, if such a study were conducted to acceptable standards, and if its results were positive and consistent, the only item lacking in comparison to two "quasi-independent" studies is a second p-value below the critical level of significance. Why is this second statistically significant result essential in establishing replicability?

Returning to the issue of demonstrating substantial evidence of efficacy, a consensus of the views expressed above can be approached by expressing more clearly the practical requirements of replicability and independence. If this can be achieved, an assessment can be begun of the degree to which a single multicenter study can be seen as satisfying (or failing to satisfy) these requirements, and the conditions required for such a study to be accepted can be discussed.

A summary of the discussion of this paper will be published in the next issue of the *Drug Information Journal*.

REFERENCES

1. Food and Drug Administration. "Guidelines for the Format and Content of the Clinical and Statistical Sections of New Drug Applications." Legislation, Professional and Consumer Affairs, Rockville, Maryland, 1988.
2. Troendle GJ. Preparing the clinical section of a new drug application. *Reg Aff.* 1989; 1:49—54.
3. Nevius SE. Assessment of evidence from a single multicenter trial. *Am Stat Assoc Biopharmaceutical Section.* 1988;43—45.