Guidance for Industry

E9 Statistical Principles for Clinical Trials

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
September 1998
ICH

Guidance for Industry

E9 Statistical Principles for Clinical Trials

Additional copies are available from:

Office of Training and Communications
Division of Drug Information (HFD-240)
Center for Drug Evaluation and Research (CDER),
5600 Fishers Lane, Rockville, MD 20857 (Tel) 301-827-4573
http://www.fda.gov/cder/guidance/index.htm

or

Office of Communication, Training, and Manufacturers Assistance (HFM-40)
Center for Biologics Evaluation and Research (CBER)
1401 Rockville Pike, Rockville, MD 20852-1448
http://www.fda.gov/cber/guidelines.htm; (Fax) 888-CBERFAX or 301-827-3844
(Voice Information) 800-835-4709 or 301-827-1800

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
September 1998
ICH

TABLE OF CONTENTS

GUIDANCE FOR INDUSTRY[1]

E9 Statistical Principles for Clinical Trials

I. INTRODUCTION

A. Background and Purpose (1.1) [2]

The efficacy and safety of medicinal products should be demonstrated by clinical trials that follow the guidance in *E6 Good Clinical Practice: Consolidated Guidance* adopted by the ICH, May 1, 1996. The role of statistics in clinical trial design and analysis is acknowledged as essential in that ICH guidance. The proliferation of statistical research in the area of clinical trials coupled with the critical role of clinical research in the drug approval process and health care in general necessitate a succinct document on statistical issues related to clinical trials. This guidance is written primarily to attempt to harmonize the principles of statistical methodology applied to clinical trials for marketing applications submitted in Europe, Japan and the United States.

As a starting point, this guidance utilized the CPMP (Committee for Proprietary Medicinal Products) Note for Guidance entitled *Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products* (December 1994). It was also influenced by *Guidelines on the Statistical Analysis of Clinical Studies* (March 1992) from the Japanese Ministry of Health and Welfare and the U.S. Food and Drug Administration document entitled *Guideline for the Format and Content of the Clinical*

[1] This guidance was developed within the Expert Working Group (Efficacy) of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) and has been subject to consultation by the regulatory parties, in accordance with the ICH process. This document has been endorsed by the ICH Steering Committee at *Step 4* of the ICH process, February 1998. At *Step 4* of the process, the final draft is recommended for adoption to the regulatory bodies of the European Union, Japan, and the United States. This guidance was published in the *Federal Register* on September 16, 1998 (63 FR 49583), and is applicable to drug and biological products.

[2] Arabic numbers reflect the organizational breakdown in the document endorsed by the ICH Steering Committee at Step 4 of the ICH process, February 1998.

and Statistical Sections of a New Drug Application (July 1988). Some topics related to statistical principles and methodology are also embedded within other ICH guidances, particularly those listed below. The specific guidance that contains related text will be identified in various sections of this document.

E1A *The Extent of Population Exposure to Assess Clinical Safety (March 1995)*
E2A *Clinical Safety Data Management: Definitions and Standards for Expedited Reporting (March 1995)*
E2B *Clinical Safety Data Management: Data Elements for Transmission of Individual Case Safety Reports (January 1998)*
E2C *Clinical Safety Data Management: Periodic Safety Update Reports for Marketed Drugs (November 1996)*
E3 *Structure and Content of Clinical Study Reports (July 1996)*
E4 *Dose-Response Information to Support Drug Registration (November 1994)*
E5 *Ethnic Factors in the Acceptability of Foreign Clinical Data (June 1998)*
E6 *Good Clinical Practice: Consolidated Guideline (April 1996)*
E7 *Studies in Support of Special Populations: Geriatrics (August 1994)*
E8 *General Considerations for Clinical Trials (December 1997*
E10 *Choice of Control Group in Clinical Trials (September 1999)*
M1 *Standardization of Medical Terminology for Regulatory Purposes (November 1999)*
M3 *Nonclinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals (July 1997)*

This guidance is intended to give direction to sponsors in the design, conduct, analysis, and evaluation of clinical trials of an investigational product in the context of its overall clinical development. The document will also assist scientific experts charged with preparing application summaries or assessing evidence of efficacy and safety, principally from clinical trials in later phases of development.

B. Scope and Direction (1.2)

The focus of this guidance is on statistical principles. It does not address the use of specific statistical procedures or methods. Specific procedural steps to ensure that principles are implemented properly are the responsibility of the sponsor. Integration of data across clinical trials is discussed, but is not a primary focus of this guidance. Selected principles and procedures related to data management or clinical trial monitoring activities are covered in other ICH guidances and are not addressed here.

This guidance should be of interest to individuals from a broad range of scientific disciplines. However, it is assumed that the actual responsibility for all statistical work associated with clinical trials will lie with an appropriately qualified and experienced statistician, as indicated in ICH E6. The role and responsibility of the trial statistician (see Glossary), in collaboration with other clinical trial professionals, is to ensure that statistical principles are applied appropriately in clinical trials supporting drug

development. Thus, the trial statistician should have a combination of education/training and experience sufficient to implement the principles articulated in this guidance.

For each clinical trial contributing to a marketing application, all important details of its design and conduct and the principal features of its proposed statistical analysis should be clearly specified in a protocol written before the trial begins. The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial. The protocol and subsequent amendments should be approved by the responsible personnel, including the trial statistician. The trial statistician should ensure that the protocol and any amendments cover all relevant statistical issues clearly and accurately, using technical terminology as appropriate.

The principles outlined in this guidance are primarily relevant to clinical trials conducted in the later phases of development, many of which are confirmatory trials of efficacy. In addition to efficacy, confirmatory trials may have as their primary variable a safety variable (e.g., an adverse event, a clinical laboratory variable, or an electrocardiographic measure) or a pharmacodynamic or pharmacokinetic variable (as in a confirmatory bioequivalence trial). Furthermore, some confirmatory findings may be derived from data integrated across trials, and selected principles in this guidance are applicable in this situation. Finally, although the early phases of drug development consist mainly of clinical trials that are exploratory in nature, statistical principles are also relevant to these clinical trials. Hence, the substance of this document should be applied as far as possible to all phases of clinical development.

Many of the principles delineated in this guidance deal with minimizing bias (see Glossary) and maximizing precision. As used in this guidance, the term *bias* describes the systematic tendency of any factors associated with the design, conduct, analysis, and interpretation of the results of clinical trials to make the estimate of a treatment effect (see Glossary) deviate from its true value. It is important to identify potential sources of bias as completely as possible so that attempts to limit such bias may be made. The presence of bias may seriously compromise the ability to draw valid conclusions from clinical trials.

Some sources of bias arise from the design of the trial, for example an assignment of treatments such that subjects at lower risk are systematically assigned to one treatment. Other sources of bias arise during the conduct and analysis of a clinical trial. For example, protocol violations and exclusion of subjects from analysis based upon knowledge of subject outcomes are possible sources of bias that may affect the accurate assessment of the treatment effect. Because bias can occur in subtle or unknown ways and its effect is not measurable directly, it is important to evaluate the robustness of the results and primary conclusions of the trial. Robustness is a concept that refers to the sensitivity of the overall conclusions to various limitations of the data, assumptions, and analytic approaches to data analysis. Robustness implies that the treatment effect and primary conclusions of the trial are not substantially affected when analyses are carried out based on alternative assumptions or analytic approaches. The interpretation of statistical measures of

uncertainty of the treatment effect and treatment comparisons should involve consideration of the potential contribution of bias to the p-value, confidence interval, or inference.

Because the predominant approaches to the design and analysis of clinical trials have been based on frequentist statistical methods, the guidance largely refers to the use of frequentist methods (see Glossary) when discussing hypothesis testing and/or confidence intervals. This should not be taken to imply that other approaches are not appropriate; the use of Bayesian (see Glossary) and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust.

II. CONSIDERATIONS FOR OVERALL CLINICAL DEVELOPMENT

A. Trial Context (2.1)

1. Development Plan (2.1.1)

The broad aim of the process of clinical development of a new drug is to find out whether there is a dose range and schedule at which the drug can be shown to be simultaneously safe and effective, to the extent that the risk-benefit relationship is acceptable. The particular subjects who may benefit from the drug, and the specific indications for its use, also need to be defined.

Satisfying these broad aims usually requires an ordered program of clinical trials, each with its own specific objectives (see ICH E8). This should be specified in a clinical plan, or a series of plans, with appropriate decision points and flexibility to allow modification as knowledge accumulates. A marketing application should clearly describe the main content of such plans, and the contribution made by each trial. Interpretation and assessment of the evidence from the total program of trials involves synthesis of the evidence from the individual trials (see section VII.B). This is facilitated by ensuring that common standards are adopted for a number of features of the trials, such as dictionaries of medical terms, definition and timing of the main measurements, handling of protocol deviations, and so on. A statistical summary, overview, or meta-analysis (see Glossary) may be informative when medical questions are addressed in more than one trial. Where possible, this should be envisaged in the plan so that the relevant trials are clearly identified and any necessary common features of their designs are specified in advance. Other major statistical issues (if any) that are expected to affect a number of trials in a common plan should be addressed in that plan.

2. Confirmatory Trial (2.1.2)

A confirmatory trial is an adequately controlled trial in which the hypotheses are stated in advance and evaluated. As a rule, confirmatory trials are necessary to provide firm evidence of efficacy or safety. In such trials the key hypothesis of interest follows directly from the trial's primary objective, is always predefined,

and is the hypothesis that is subsequently tested when the trial is complete. In a confirmatory trial, it is equally important to estimate with due precision the size of the effects attributable to the treatment of interest and to relate these effects to their clinical significance.

Confirmatory trials are intended to provide firm evidence in support of claims; hence adherence to protocols and standard operating procedures is particularly important. Unavoidable changes should be explained and documented, and their effect examined. A justification of the design of each such trial and of other important statistical aspects, such as the principal features of the planned analysis, should be set out in the protocol. Each trial should address only a limited number of questions.

Firm evidence in support of claims requires that the results of the confirmatory trials demonstrate that the investigational product under test has clinical benefits. The confirmatory trials should therefore be sufficient to answer each key clinical question relevant to the efficacy or safety claim clearly and definitively. In addition, it is important that the basis for generalization (see Glossary) to the intended patient population is understood and explained; this may also influence the number and type (e.g., specialist or general practitioner) of centers and/or trials needed. The results of the confirmatory trial(s) should be robust. In some circumstances, the weight of evidence from a single confirmatory trial may be sufficient.

3. Exploratory Trial (2.1.3)

The rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out in a series of exploratory studies. Like all clinical trials, these exploratory studies should have clear and precise objectives. However, in contrast to confirmatory trials, their objectives may not always lead to simple tests of predefined hypotheses. In addition, exploratory trials may sometimes require a more flexible approach to design so that changes can be made in response to accumulating results. Their analysis may entail data exploration. Tests of hypothesis may be carried out, but the choice of hypothesis may be data dependent. Such trials cannot be the basis of the formal proof of efficacy, although they may contribute to the total body of relevant evidence.

Any individual trial may have both confirmatory and exploratory aspects. For example, in most confirmatory trials the data are also subjected to exploratory analyses which serve as a basis for explaining or supporting their findings and for suggesting further hypotheses for later research. The protocol should make a clear distinction between the aspects of a trial which will be used for confirmatory proof and the aspects which will provide data for exploratory analysis.

B. Scope of Trials (2.2)

1. Population (2.2.1)

In the earlier phases of drug development, the choice of subjects for a clinical trial may be heavily influenced by the wish to maximize the chance of observing specific clinical effects of interest. Hence they may come from a very narrow subgroup of the total patient population for which the drug may eventually be indicated. However, by the time the confirmatory trials are undertaken, the subjects in the trials should more closely mirror the target population. In these trials, it is generally helpful to relax the inclusion and exclusion criteria as much as possible within the target population while maintaining sufficient homogeneity to permit precise estimation of treatment effects. No individual clinical trial can be expected to be totally representative of future users because of the possible influences of geographical location, the time when it is conducted, the medical practices of the particular investigator(s) and clinics, and so on. However, the influence of such factors should be reduced wherever possible and subsequently discussed during the interpretation of the trial results.

2. Primary and Secondary Variables (2.2.2)

The primary variable (*target* variable, primary endpoint) should be the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial. There should generally be only one primary variable. This will usually be an efficacy variable, because the primary objective of most confirmatory trials is to provide strong scientific evidence regarding efficacy. Safety/tolerability may sometimes be the primary variable, and will always be an important consideration. Measurements relating to quality of life and health economics are further potential primary variables. The selection of the primary variable should reflect the accepted norms and standards in the relevant field of research. The use of a reliable and validated variable with which experience has been gained either in earlier studies or in published literature is recommended. There should be sufficient evidence that the primary variable can provide a valid and reliable measure of some clinically relevant and important treatment benefit in the patient population described by the inclusion and exclusion criteria. The primary variable should generally be the one used when estimating the sample size (see section III.E).

In many cases, the approach to assessing subject outcome may not be straightforward and should be carefully defined. For example, it is inadequate to specify mortality as a primary variable without further clarification; mortality may be assessed by comparing proportions alive at fixed points in time or by comparing overall distributions of survival times over a specified interval. Another common example is a recurring event; the measure of treatment effect may again be a simple dichotomous variable (any occurrence during a specified interval), time to first

occurrence, rate of occurrence (events per time units of observation), and so on. The assessment of functional status over time in studying treatment for chronic disease presents other challenges in selection of the primary variable. There are many possible approaches, such as comparisons of the assessments done at the beginning and end of the interval of observation, comparisons of slopes calculated from all assessments throughout the interval, comparisons of the proportions of subjects exceeding or declining beyond a specified threshold**,** or comparisons based on methods for repeated measures data. To avoid multiplicity concerns arising from post hoc definitions, it is critical to specify in the protocol the precise definition of the primary variable as it will be used in the statistical analysis. In addition, the clinical relevance of the specific primary variable selected and the validity of the associated measurement procedures will generally need to be addressed and justified in the protocol.

The primary variable should be specified in the protocol, along with the rationale for its selection. Redefinition of the primary variable after unblinding will almost always be unacceptable, since the biases this introduces are difficult to assess. When the clinical effect defined by the primary objective is to be measured in more than one way, the protocol should identify one of the measurements as the primary variable on the basis of clinical relevance, importance, objectivity, and/or other relevant characteristics, whenever such selection is feasible.

Secondary variables are either supportive measurements related to the primary objective or measurements of effects related to the secondary objectives. Their predefinition in the protocol is also important, as well as an explanation of their relative importance and roles in interpretation of trial results. The number of secondary variables should be limited and should be related to the limited number of questions to be answered in the trial.

3. Composite Variables (2.2.3)

If a single primary variable cannot be selected from multiple measurements associated with the primary objective, another useful strategy is to integrate or combine the multiple measurements into a single or *composite* variable, using a predefined algorithm. Indeed, the primary variable sometimes arises as a combination of multiple clinical measurements (e.g., the rating scales used in arthritis, psychiatric disorders, and elsewhere). This approach addresses the multiplicity problem without requiring adjustment to the Type I error. The method of combining the multiple measurements should be specified in the protocol, and an interpretation of the resulting scale should be provided in terms of the size of a clinically relevant benefit. When a composite variable is used as a primary variable, the components of this variable may sometimes be analyzed separately, where clinically meaningful and validated. When a rating scale is used as a primary variable, it is especially important to address factors such as content

validity (see Glossary), inter- and intrarater reliability (see Glossary), and responsiveness for detecting changes in the severity of disease.

4. Global Assessment Variables (2.2.4)

In some cases, *global assessment* variables (see Glossary) are developed to measure the overall safety, overall efficacy, and/or overall usefulness of a treatment. This type of variable integrates objective variables and the investigator's overall impression about the state or change in the state of the subject, and is usually a scale of ordered categorical ratings. Global assessments of overall efficacy are well established in some therapeutic areas, such as neurology and psychiatry.

Global assessment variables generally have a subjective component. When a global assessment variable is used as a primary or secondary variable, fuller details of the scale should be included in the protocol with respect to:

- The relevance of the scale to the primary objective of the trial;
- The basis for the validity and reliability of the scale;
- How to utilize the data collected on an individual subject to assign him/her to a unique category of the scale;
- How to assign subjects with missing data to a unique category of the scale, or otherwise evaluate them.

If objective variables are considered by the investigator when making a global assessment, then those objective variables should be considered as additional primary or, at least, important secondary variables.

Global assessment of usefulness integrates components of both benefit and risk and reflects the decisionmaking process of the treating physician, who must weigh benefit and risk in making product use decisions. A problem with global usefulness variables is that their use could in some cases lead to the result of two products being declared equivalent despite having very different profiles of beneficial and adverse effects. For example, judging the global usefulness of a treatment as equivalent or superior to an alternative may mask the fact that it has little or no efficacy but fewer adverse effects. Therefore, it is not advisable to use a global usefulness variable as a primary variable. If global usefulness is specified as primary, it is important to consider specific efficacy and safety outcomes separately as additional primary variables.

5. Multiple Primary Variables (2.2.5)

It may sometimes be desirable to use more than one primary variable, each of which (or a subset of which) could be sufficient to cover the range of effects of the therapies. The planned manner of interpretation of this type of evidence should be carefully spelled out. It should be clear whether an impact on any of the variables, some minimum number of them, or all of them, would be considered necessary to achieve the trial objectives. The primary hypothesis or hypotheses and parameters of interest (e.g., mean, percentage, distribution) should be clearly stated with respect to the primary variables identified, and the approach to statistical inference described. The effect on the Type I error should be explained because of the potential for multiplicity problems (see section V.F); the method of controlling Type I error should be given in the protocol. The extent of intercorrelation among the proposed primary variables may be considered in evaluating the impact on Type I error. If the purpose of the trial is to demonstrate effects on all of the designated primary variables, then there is no need for adjustment of the Type I error, but the impact on Type II error and sample size should be carefully considered.

6. Surrogate Variables (2.2.6)

When direct assessment of the clinical benefit to the subject through observing actual clinical efficacy is not practical, indirect criteria (surrogate variables — see Glossary) may be considered. Commonly accepted surrogate variables are used in a number of indications where they are believed to be reliable predictors of clinical benefit. There are two principal concerns with the introduction of any proposed surrogate variable. First, it may not be a true predictor of the clinical outcome of interest. For example, it may measure treatment activity associated with one specific pharmacological mechanism, but may not provide full information on the range of actions and ultimate effects of the treatment, whether positive or negative. There have been many instances where treatments showing a highly positive effect on a proposed surrogate have ultimately been shown to be detrimental to the subjects' clinical outcome; conversely, there are cases of treatments conferring clinical benefit without measurable impact on proposed surrogates. Second, proposed surrogate variables may not yield a quantitative measure of clinical benefit that can be weighed directly against adverse effects. Statistical criteria for validating surrogate variables have been proposed but the experience with their use is relatively limited. In practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome, and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome. Relationships between clinical and surrogate variables for one product do not necessarily apply to a product with a different mode of action for treating the same disease.

7. Categorized Variables (2.2.7)

Dichotomization or other categorization of continuous or ordinal variables may sometimes be desirable. Criteria of *success* and *response* are common examples of dichotomies that should be specified precisely in terms of, for example, a minimum percentage improvement (relative to baseline) in a continuous variable or a ranking categorized as at or above some threshold level (e.g., *good*) on an ordinal rating scale. The reduction of diastolic blood pressure below 90 mmHg is a common dichotomization. Categorizations are most useful when they have clear clinical relevance. The criteria for categorization should be predefined and specified in the protocol, as knowledge of trial results could easily bias the choice of such criteria. Because categorization normally implies a loss of information, a consequence will be a loss of power in the analysis; this should be accounted for in the sample size calculation.

C. Design Techniques to Avoid Bias (2.3)

The most important design techniques for avoiding bias in clinical trials are blinding and randomization, and these should be normal features of most controlled clinical trials intended to be included in a marketing application. Most such trials follow a double-blind approach in which treatments are prepacked in accordance with a suitable randomization schedule, and supplied to the trial center(s) labeled only with the subject number and the treatment period, so that no one involved in the conduct of the trial is aware of the specific treatment allocated to any particular subject, not even as a code letter. This approach will be assumed in section II.C.1 and most of section II.C.2, exceptions being considered at the end.

Bias can also be reduced at the design stage by specifying procedures in the protocol aimed at minimizing any anticipated irregularities in trial conduct that might impair a satisfactory analysis, including various types of protocol violations, withdrawals and missing values. The protocol should consider ways both to reduce the frequency of such problems and to handle the problems that do occur in the analysis of data.

1. Blinding (2.3.1)

Blinding or masking is intended to limit the occurrence of conscious and unconscious bias in the conduct and interpretation of a clinical trial arising from the influence that the knowledge of treatment may have on the recruitment and allocation of subjects, their subsequent care, the attitudes of subjects to the treatments, the assessment of end-points, the handling of withdrawals, the exclusion of data from analysis, and so on. The essential aim is to prevent identification of the treatments until all such opportunities for bias have passed.

A double-blind trial is one in which neither the subject nor any of the investigator or sponsor staff involved in the treatment or clinical evaluation of the subjects are aware of the treatment received. This includes anyone determining subject eligibility, evaluating endpoints, or assessing compliance with the protocol. This level of blinding is maintained throughout the conduct of the trial, and only when the data are cleaned to an acceptable level of quality will appropriate personnel be unblinded. If any of the sponsor staff who are not involved in the treatment or clinical evaluation of the subjects are required to be unblinded to the treatment code (e.g., bioanalytical scientists, auditors, those involved in serious adverse event reporting), the sponsor should have adequate standard operating procedures to guard against inappropriate dissemination of treatment codes. In a single-blind trial the investigator and/or his staff are aware of the treatment but the subject is not, or vice versa. In an open-label trial the identity of treatment is known to all. The double-blind trial is the optimal approach. This requires that the treatments to be applied during the trial cannot be distinguished (by appearance, taste, etc.) either before or during administration, and that the blind is maintained appropriately during the whole trial.

Difficulties in achieving the double-blind ideal can arise: The treatments may be of a completely different nature, for example, surgery and drug therapy; two drugs may have different formulations and, although they could be made indistinguishable by the use of capsules, changing the formulation might also change the pharmacokinetic and/or pharmacodynamic properties and hence necessitate that bioequivalence of the formulations be established; the daily pattern of administration of two treatments may differ. One way of achieving double-blind conditions under these circumstances is to use a *double-dummy* (see Glossary) technique. This technique may sometimes force an administration scheme that is sufficiently unusual to influence adversely the motivation and compliance of the subjects. Ethical difficulties may also interfere with its use when, for example, it entails dummy operative procedures. Nevertheless, extensive efforts should be made to overcome these difficulties.

The double-blind nature of some clinical trials may be partially compromised by apparent treatment induced effects. In such cases, blinding may be improved by blinding investigators and relevant sponsor staff to certain test results (e.g., selected clinical laboratory measures). Similar approaches (see below) to minimizing bias in open-label trials should be considered in trials where unique or specific treatment effects may lead to unblinding individual patients.

If a double-blind trial is not feasible, then the single-blind option should be considered. In some cases only an open-label trial is practically or ethically possible. Single-blind and open-label trials provide additional flexibility, but it is particularly important that the investigator's knowledge of the next treatment should not influence the decision to enter the subject; this decision should precede knowledge of the randomized treatment. For these trials, consideration should be given to the use of a centralized randomization method, such as telephone

randomization, to administer the assignment of randomized treatment. In addition, clinical assessments should be made by medical staff who are not involved in treating the subjects and who remain blind to treatment. In single-blind or open-label trials every effort should be made to minimize the various known sources of bias and primary variables should be as objective as possible. The reasons for the degree of blinding adopted, as well as steps taken to minimize bias by other means, should be explained in the protocol. For example, the sponsor should have adequate standard operating procedures to ensure that access to the treatment code is appropriately restricted during the process of cleaning the database prior to its release for analysis.

Breaking the blind (for a single subject) should be considered only when knowledge of the treatment assignment is deemed essential by the subject's physician for the subject's care. Any intentional or unintentional breaking of the blind should be reported and explained at the end of the trial, irrespective of the reason for its occurrence. The procedure and timing for revealing the treatment assignments should be documented.

In this document, the blind review (see Glossary) of data refers to the checking of data during the period of time between trial completion (the last observation on the last subject) and the breaking of the blind.

2. Randomization (2.3.2)

Randomization introduces a deliberate element of chance into the assignment of treatments to subjects in a clinical trial. During subsequent analysis of the trial data, it provides a sound statistical basis for the quantitative evaluation of the evidence relating to treatment effects. It also tends to produce treatment groups in which the distributions of prognostic factors, known and unknown, are similar. In combination with blinding, randomization helps to avoid possible bias in the selection and allocation of subjects arising from the predictability of treatment assignments.

The randomization schedule of a clinical trial documents the random allocation of treatments to subjects. In the simplest situation it is a sequential list of treatments (or treatment sequences in a crossover trial) or corresponding codes by subject number. The logistics of some trials, such as those with a screening phase, may make matters more complicated, but the unique preplanned assignment of treatment, or treatment sequence, to subject should be clear. Different trial designs will necessitate different procedures for generating randomization schedules. The randomization schedule should be reproducible (if the need arises).

Although unrestricted randomization is an acceptable approach, some advantages can generally be gained by randomizing subjects in blocks. This helps to increase the comparability of the treatment groups, particularly when subject characteristics may change over time, as a result, for example, of changes in recruitment policy. It also provides a better guarantee that the treatment groups will be of nearly equal

size. In crossover trials, it provides the means of obtaining balanced designs with their greater efficiency and easier interpretation. Care should be taken to choose block lengths that are sufficiently short to limit possible imbalance, but that are long enough to avoid predictability towards the end of the sequence in a block. Investigators and other relevant staff should generally be blind to the block length; the use of two or more block lengths, randomly selected for each block, can achieve the same purpose. (Theoretically, in a double-blind trial predictability does not matter, but the pharmacological effects of drugs may provide the opportunity for intelligent guesswork.)

In multicenter trials (see Glossary), the randomization procedures should be organized centrally. It is advisable to have a separate random scheme for each center, i.e., to stratify by center or to allocate several whole blocks to each center. More generally, stratification by important prognostic factors measured at baseline (e.g., severity of disease, age, sex) may sometimes be valuable in order to promote balanced allocation within strata; this has greater potential benefit in small trials. The use of more than two or three stratification factors is rarely necessary, is less successful at achieving balance, and is logistically troublesome. The use of a dynamic allocation procedure (see below) may help to achieve balance across a number of stratification factors simultaneously, provided the rest of the trial procedures can be adjusted to accommodate an approach of this type. Factors on which randomization has been stratified should be accounted for later in the analysis.

The next subject to be randomized into a trial should always receive the treatment corresponding to the next free number in the appropriate randomization schedule (in the respective stratum, if randomization is stratified). The appropriate number and associated treatment for the next subject should only be allocated when entry of that subject to the randomized part of the trial has been confirmed. Details of the randomization that facilitate predictability (e.g., block length) should not be contained in the trial protocol. The randomization schedule itself should be filed securely by the sponsor or an independent party in a manner that ensures that blindness is properly maintained throughout the trial. Access to the randomization schedule during the trial should take into account the possibility that, in an emergency, the blind may have to be broken for any subject. The procedure to be followed, the necessary documentation, and the subsequent treatment and assessment of the subject should all be described in the protocol.

Dynamic allocation is an alternative procedure in which the allocation of treatment to a subject is influenced by the current balance of allocated treatments and, in a stratified trial, by the stratum to which the subject belongs and the balance within that stratum. Deterministic dynamic allocation procedures should be avoided and an appropriate element of randomization should be incorporated for each treatment allocation. Every effort should be made to retain the double-blind status of the trial. For example, knowledge of the treatment code may be restricted to a central trial office from where the dynamic allocation is controlled, generally through telephone

contact. This in turn permits additional checks of eligibility criteria and establishes entry into the trial, features that can be valuable in certain types of multicenter trials. The usual system of prepacking and labeling drug supplies for double-blind trials can then be followed, but the order of their use is no longer sequential. It is desirable to use appropriate computer algorithms to keep personnel at the central trial office blind to the treatment code. The complexity of the logistics and potential impact on the analysis should be carefully evaluated when considering dynamic allocation.

III. TRIAL DESIGN CONSIDERATIONS

A. Design Configuration (3.1)

1. Parallel Group Design (3.1.1)

The most common clinical trial design for confirmatory trials is the parallel group design in which subjects are randomized to one of two or more arms, each arm being allocated a different treatment. These treatments will include the investigational product at one or more doses, and one or more control treatments, such as placebo and/or an active comparator. The assumptions underlying this design are less complex than for most other designs. However, as with other designs, there may be additional features of the trial that complicate the analysis and interpretation (e.g., covariates, repeated measurements over time, interactions between design factors, protocol violations, dropouts (see Glossary), and withdrawals).

2. Crossover Design (3.1.2)

In the crossover design, each subject is randomized to a sequence of two or more treatments and hence acts as his own control for treatment comparisons. This simple maneuver is attractive primarily because it reduces the number of subjects and usually the number of assessments needed to achieve a specific power, sometimes to a marked extent. In the simplest 2x2 crossover design, each subject receives each of two treatments in randomized order in two successive treatment periods, often separated by a washout period. The most common extension of this entails comparing n(>2) treatments in n periods, each subject receiving all n treatments. Numerous variations exist, such as designs in which each subject receives a subset of n(>2) treatments, or designs in which treatments are repeated within a subject.

Crossover designs have a number of problems that can invalidate their results. The chief difficulty concerns carryover, that is, the residual influence of treatments in subsequent treatment periods. In an additive model, the effect of unequal carryover will be to bias direct treatment comparisons. In the 2x2 design, the carryover effect cannot be statistically distinguished from the interaction between treatment and

period and the test for either of these effects lacks power because the corresponding contrast is *between subject*. This problem is less acute in higher order designs, but cannot be entirely dismissed.

When the crossover design is used, it is therefore important to avoid carryover. This is best done by selective and careful use of the design on the basis of adequate knowledge of both the disease area and the new medication. The disease under study should be chronic and stable. The relevant effects of the medication should develop fully within the treatment period. The washout periods should be sufficiently long for complete reversibility of drug effect. The fact that these conditions are likely to be met should be established in advance of the trial by means of prior information and data.

There are additional problems that need careful attention in crossover trials. The most notable of these are the complications of analysis and interpretation arising from the loss of subjects. Also, the potential for carryover leads to difficulties in assigning adverse events that occur in later treatment periods to the appropriate treatment. These and other issues are described in ICH E4. The crossover design should generally be restricted to situations where losses of subjects from the trial are expected to be small.

A common, and generally satisfactory, use of the 2x2 crossover design is to demonstrate the bioequivalence of two formulations of the same medication. In this particular application in healthy volunteers, carryover effects on the relevant pharmacokinetic variable are most unlikely to occur if the wash-out time between the two periods is sufficiently long. However, it is still important to check this assumption during analysis on the basis of the data obtained, for example, by demonstrating that no drug is detectable at the start of each period.

3. Factorial Designs (3.1.3)

In a factorial design, two or more treatments are evaluated simultaneously through the use of varying combinations of the treatments. The simplest example is the 2x2 factorial design in which subjects are randomly allocated to one of the four possible combinations of two treatments, A and B. These are: A alone; B alone; both A and B; neither A nor B. In many cases, this design is used for the specific purpose of examining the interaction of A and B. The statistical test of interaction may lack power to detect an interaction if the sample size was calculated based on the test for main effects. This consideration is important when this design is used for examining the joint effects of A and B, in particular, if the treatments are likely to be used together.

Another important use of the factorial design is to establish the dose-response characteristics of the simultaneous use of treatments C and D, especially when the efficacy of each monotherapy has been established at some dose in prior trials. A number, m, of doses of C is selected, usually including a zero dose (placebo), and a

similar number, n, of doses of D. The full design then consists of mxn treatment groups, each receiving a different combination of doses of C and D. The resulting estimate of the response surface may then be used to help identify an appropriate combination of doses of C and D for clinical use (see ICH E4).

In some cases, the 2x2 design may be used to make efficient use of clinical trial subjects by evaluating the efficacy of the two treatments with the same number of subjects as would be required to evaluate the efficacy of either one alone. This strategy has proved to be particularly valuable for very large mortality trials. The efficiency and validity of this approach depends upon the absence of interaction between treatments A and B so that the effects of A and B on the primary efficacy variables follow an additive model. Hence the effect of A is virtually identical whether or not it is additional to the effect of B. As for the crossover trial, evidence that this condition is likely to be met should be established in advance of the trial by means of prior information and data.

B. Multicenter Trials (3.2)

Multicenter trials are carried out for two main reasons. First, a multicenter trial is an accepted way of evaluating a new medication more efficiently. Under some circumstances, it may present the only practical means of accruing sufficient subjects to satisfy the trial objective within a reasonable timeframe. Multicenter trials of this nature may, in principle, be carried out at any stage of clinical development. They may have several centers with a large number of subjects per center or, in the case of a rare disease, they may have a large number of centers with very few subjects per center.

Second, a trial may be designed as a multicenter (and multi-investigator) trial primarily to provide a better basis for the subsequent generalization of its findings. This arises from the possibility of recruiting the subjects from a wider population and of administering the medication in a broader range of clinical settings, thus presenting an experimental situation that is more typical of future use. In this case, the involvement of a number of investigators also gives the potential for a wider range of clinical judgement concerning the value of the medication. Such a trial would be a confirmatory trial in the later phases of drug development and would be likely to involve a large number of investigators and centers. It might sometimes be conducted in a number of different countries to facilitate generalizability (see Glossary) even further.

If a multicenter trial is to be meaningfully interpreted and extrapolated, then the manner in which the protocol is implemented should be clear and similar at all centers. Furthermore, the usual sample size and power calculations depend upon the assumption that the differences between the compared treatments in the centers are unbiased estimates of the same quantity. It is important to design the common protocol and to conduct the trial with this background in mind. Procedures should be standardized as completely as possible. Variation of evaluation criteria and schemes can be reduced by investigator meetings, by the training of personnel in advance of the trial, and by careful monitoring during the trial.

Good design should generally aim to achieve the same distribution of subjects to treatments within each center and good management should maintain this design objective. Trials that avoid excessive variation in the numbers of subjects per center and trials that avoid a few very small centers have advantages if it is later found necessary to take into account the heterogeneity of the treatment effect from center to center, because they reduce the differences between different weighted estimates of the treatment effect. (This point does not apply to trials in which all centers are very small and in which center does not feature in the analysis.) Failure to take these precautions, combined with doubts about the homogeneity of the results, may, in severe cases, reduce the value of a multicenter trial to such a degree that it cannot be regarded as giving convincing evidence for the sponsor's claims.

In the simplest multicenter trial, each investigator will be responsible for the subjects recruited at one hospital, so that *center* is identified uniquely by either investigator or hospital. In many trials, however, the situation is more complex. One investigator may recruit subjects from several hospitals; one investigator may represent a team of clinicians (subinvestigators) who all recruit subjects from their own clinics at one hospital or at several associated hospitals. Whenever there is room for doubt about the definition of center in a statistical model, the statistical section of the protocol (see section V.A) should clearly define the term (e.g., by investigator, location or region) in the context of the particular trial. In most instances, centers can be satisfactorily defined through the investigators. (ICH E6 provides relevant guidance in this respect.) In cases of doubt, the aim should be to define centers to achieve homogeneity in the important factors affecting the measurements of the primary variables and the influence of the treatments. Any rules for combining centers in the analysis should be justified and specified prospectively in the protocol where possible, but in any case decisions concerning this approach should always be taken blind to treatment, for example, at the time of the blind review.

The statistical model to be adopted for the estimation and testing of treatment effects should be described in the protocol. The main treatment effect may be investigated first using a model that allows for center differences, but does not include a term for treatment-by-center interaction. If the treatment effect is homogeneous across centers, the routine inclusion of interaction terms in the model reduces the efficiency of the test for the main effects. In the presence of true heterogeneity of treatment effects, the interpretation of the main treatment effect is controversial.

In some trials, for example, some large mortality trials with very few subjects per center, there may be no reason to expect the centers to have any influence on the primary or secondary variables because they are unlikely to represent influences of clinical importance. In other trials, it may be recognized from the start that the limited numbers of subjects per center will make it impracticable to include the center effects in the statistical model. In these cases, it is not considered appropriate to include a term for center in the model, and it is not necessary to stratify the randomization by center in this situation.

If positive treatment effects are found in a trial with appreciable numbers of subjects per center, there should generally be an exploration of the heterogeneity of treatment effects

across centers, as this may affect the generalizability of the conclusions. Marked heterogeneity may be identified by graphical display of the results of individual centers or by analytical methods, such as a significance test of the treatment-by-center interaction. When using such a statistical significance test, it is important to recognize that this generally has low power in a trial designed to detect the main effect of treatment.

If heterogeneity of treatment effects is found, this should be interpreted with care, and vigorous attempts should be made to find an explanation in terms of other features of trial management or subject characteristics. Such an explanation will usually suggest appropriate further analysis and interpretation. In the absence of an explanation, heterogeneity of treatment effect, as evidenced, for example, by marked quantitative interactions (see Glossary) implies that alternative estimates of the treatment effect, giving different weights to the centers, may be needed to substantiate the robustness of the estimates of treatment effect. It is even more important to understand the basis of any heterogeneity characterized by marked qualitative interactions (see Glossary), and failure to find an explanation may necessitate further clinical trials before the treatment effect can be reliably predicted.

Up to this point, the discussion of multicenter trials has been based on the use of fixed effect models. Mixed models may also be used to explore the heterogeneity of the treatment effect. These models consider center and treatment-by-center effects to be random and are especially relevant when the number of sites is large.

C. Type of Comparison (3.3)

1. Trials to Show Superiority (3.3.1)

Scientifically, efficacy is most convincingly established by demonstrating superiority to placebo in a placebo-controlled trial, by showing superiority to an active control treatment, or by demonstrating a dose-response relationship. This type of trial is referred to as a *superiority* trial (see Glossary). In this guidance superiority trials are generally assumed, unless explicitly stated otherwise.

For serious illnesses, when a therapeutic treatment that has been shown to be efficacious by superiority trial(s) exists, a placebo-controlled trial may be considered unethical. In that case the scientifically sound use of an active treatment as a control should be considered. The appropriateness of placebo control versus active control should be considered on a trial-by-trial basis.

2. Trials to Show Equivalence or Noninferiority (3.3.2)

In some cases, an investigational product is compared to a reference treatment without the objective of showing superiority. This type of trial is divided into two major categories according to its objective; one is an *equivalence* trial (see Glossary) and the other is a *noninferiority* trial (see Glossary).

Bioequivalence trials fall into the former category. In some situations, clinical equivalence trials are also undertaken for other regulatory reasons such as demonstrating the clinical equivalence of a generic product to the marketed product when the compound is not absorbed and therefore not present in the blood stream.

Many active control trials are designed to show that the efficacy of an investigational product is no worse than that of the active comparator and, hence, fall into the latter category. Another possibility is a trial in which multiple doses of the investigational drug are compared with the recommended dose or multiple doses of the standard drug. The purpose of this design is simultaneously to show a dose-response relationship for the investigational product and to compare the investigational product with the active control.

Active control equivalence or noninferiority trials may also incorporate a placebo, thus pursuing multiple goals in one trial. For example, they may establish superiority to placebo and hence validate the trial design and simultaneously evaluate the degree of similarity of efficacy and safety to the active comparator. There are well-known difficulties associated with the use of the active control equivalence (or noninferiority) trials that do not incorporate a placebo or do not use multiple doses of the new drug. These relate to the implicit lack of any measure of internal validity (in contrast to superiority trials), thus making external validation necessary. The equivalence (or noninferiority) trial is not conservative in nature, so that many flaws in the design or conduct of the trial will tend to bias the results towards a conclusion of equivalence. For these reasons, the design features of such trials should receive special attention and their conduct needs special care. For example, it is especially important to minimize the incidence of violations of the entry criteria, noncompliance, withdrawals, losses to follow-up, missing data, and other deviations from the protocol, and also to minimize their impact on the subsequent analyses.

Active comparators should be chosen with care. An example of a suitable active comparator would be a widely used therapy whose efficacy in the relevant indication has been clearly established and quantified in well-designed and well-documented superiority trial(s) and that can be reliably expected to exhibit similar efficacy in the contemplated active control trial. To this end, the new trial should have the same important design features (primary variables, the dose of the active comparator, eligibility criteria, and so on) as the previously conducted superiority trials in which the active comparator clearly demonstrated clinically relevant efficacy, taking into account advances in medical or statistical practice relevant to the new trial.

It is vital that the protocol of a trial designed to demonstrate equivalence or noninferiority contain a clear statement that this is its explicit intention. An equivalence margin should be specified in the protocol; this margin is the largest difference that can be judged as being clinically acceptable and should be smaller

than differences observed in superiority trials of the active comparator. For the active control equivalence trial, both the upper and the lower equivalence margins are needed, while only the lower margin is needed for the active control noninferiority trial. The choice of equivalence margins should be justified clinically.

Statistical analysis is generally based on the use of confidence intervals (see section V.E). For equivalence trials, two-sided confidence intervals should be used. Equivalence is inferred when the entire confidence interval falls within the equivalence margins. Operationally, this is equivalent to the method of using two simultaneous one-sided tests to test the (composite) null hypothesis that the treatment difference is outside the equivalence margins versus the (composite) alternative hypothesis that the treatment difference is within the margins. Because the two null hypotheses are disjoint, the Type I error is appropriately controlled. For noninferiority trials, a one-sided interval should be used. The confidence interval approach has a one-sided hypothesis test counterpart for testing the null hypothesis that the treatment difference (investigational product minus control) is equal to the lower equivalence margin versus the alternative that the treatment difference is greater than the lower equivalence margin. The choice of Type I error should be a consideration separate from the use of a one-sided or two-sided procedure. Sample size calculations should be based on these methods (see section III.E).

Concluding equivalence or noninferiority based on observing a nonsignificant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is considered inappropriate.

There are also special issues in the choice of analysis sets. Subjects who withdraw or drop out of the treatment group or the comparator group will tend to have a lack of response; hence the results of using the full analysis set (see Glossary) may be biased toward demonstrating equivalence (see section V.B.3).

3. Trials to Show Dose-Response Relationship (3.3.3)

How response is related to the dose of a new investigational product is a question to which answers may be obtained in all phases of development and by a variety of approaches (see ICH E4). Dose-response trials may serve a number of objectives, among which the following are of particular importance: the confirmation of efficacy; the investigation of the shape and location of the dose-response curve; the estimation of an appropriate starting dose; the identification of optimal strategies for individual dose adjustments; the determination of a maximal dose beyond which additional benefit would be unlikely to occur. These objectives should be addressed using the data collected at a number of doses under investigation, including a placebo (zero dose) wherever appropriate. For this purpose, the application of procedures to estimate the relationship between dose and response,

including the construction of confidence intervals and the use of graphical methods, is as important as the use of statistical tests. The hypothesis tests that are used may need to be tailored to the natural ordering of doses or to particular questions regarding the shape of the dose-response curve (e.g., monotonicity). The details of the planned statistical procedures should be given in the protocol.

D. Group Sequential Designs (3.4)

Group sequential designs are used to facilitate the conduct of interim analysis (see section IV.E and Glossary). While group sequential designs are not the only acceptable types of designs permitting interim analysis, they are the most commonly applied because it is more practicable to assess grouped subject outcomes at periodic intervals during the trial than on a continuous basis as data from each subject become available. The statistical methods should be fully specified in advance of the availability of information on treatment outcomes and subject treatment assignments (i.e., blind breaking, see section IV.E). An independent data monitoring committee (IDMC) (see Glossary) may be used to review or to conduct the interim analysis of data arising from a group sequential design (see section IV.F). While the design has been most widely and successfully used in large, long-term trials of mortality or major nonfatal endpoints, its use is growing in other circumstances. In particular, it is recognized that safety must be monitored in all trials; therefore, the need for formal procedures to cover early stopping for safety reasons should always be considered.

E. Sample Size (3.5)

The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed. This number is usually determined by the primary objective of the trial. If the sample size is determined on some other basis, then this should be made clear and justified. For example, a trial sized on the basis of safety questions or requirements or important secondary objectives may need larger numbers of subjects than a trial sized on the basis of the primary efficacy question (see ICH E1A).

Using the usual method for determining the appropriate sample size, the following items should be specified: A primary variable; the test statistic; the null hypothesis; the alternative (*working*) hypothesis at the chosen dose(s) (embodying consideration of the treatment difference to be detected or rejected at the dose and in the subject population selected); the probability of erroneously rejecting the null hypothesis (the Type I error) and the probability of erroneously failing to reject the null hypothesis (the Type II error); as well as the approach to dealing with treatment withdrawals and protocol violations. In some instances, the event rate is of primary interest for evaluating power, and assumptions should be made to extrapolate from the required number of events to the eventual sample size for the trial.

The method by which the sample size is calculated should be given in the protocol, together with the estimates of any quantities used in the calculations (such as variances, mean values, response rates, event rates, difference to be detected). The basis of these

estimates should also be given. It is important to investigate the sensitivity of the sample size estimate to a variety of deviations from these assumptions and this may be facilitated by providing a range of sample sizes appropriate for a reasonable range of deviations from assumptions. In confirmatory trials, assumptions should normally be based on published data or on the results of earlier trials. The treatment difference to be detected may be based on a judgement concerning the minimal effect which has clinical relevance in the management of patients or on a judgement concerning the anticipated effect of the new treatment, where this is larger. Conventionally, the probability of Type I error is set at 5 percent or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of Type II error is conventionally set at 10 percent to 20 percent. It is in the sponsor's interest to keep this figure as low as feasible, especially in the case of trials that are difficult or impossible to repeat. Alternative values to the conventional levels of Type I and Type II error may be acceptable or even preferable in some cases.

Sample size calculations should refer to the number of subjects required for the primary analysis. If this is the *full analysis set*, estimates of the effect size may need to be reduced compared to the per protocol set (see Glossary). This is to allow for the dilution of the treatment effect arising from the inclusion of data from patients who have withdrawn from treatment or whose compliance is poor. The assumptions about variability may also need to be revised.

The sample size of an equivalence trial or a noninferiority trial (see section III.C.2) should normally be based on the objective of obtaining a confidence interval for the treatment difference that shows that the treatments differ at most by a clinically acceptable difference. When the power of an equivalence trial is assessed at a true difference of zero, then the sample size necessary to achieve this power is underestimated if the true difference is not zero. When the power of a noninferiority trial is assessed at a zero difference, then the sample size needed to achieve that power will be underestimated if the effect of the investigational product is less than that of the active control. The choice of a *clinically acceptable* difference needs justification with respect to its meaning for future patients, and may be smaller than the "clinically relevant" difference referred to above in the context of superiority trials designed to establish that a difference exists.

The exact sample size in a group sequential trial cannot be fixed in advance because it depends upon the play of chance in combination with the chosen stopping guideline and the true treatment difference. The design of the stopping guideline should take into account the consequent distribution of the sample size, usually embodied in the expected and maximum sample sizes.

When event rates are lower than anticipated or variability is larger than expected, methods for sample size reestimation are available without unblinding data or making treatment comparisons (see section IV.D).

F. Data Capture and Processing (3.6)

The collection of data and transfer of data from the investigator to the sponsor can take place through a variety of media, including paper case record forms, remote site monitoring systems, medical computer systems, and electronic transfer. Whatever data capture instrument is used, the form and content of the information collected should be in full accordance with the protocol and should be established in advance of the conduct of the clinical trial. It should focus on the data necessary to implement the planned analysis, including the context information (such as timing assessments relative to dosing) necessary to confirm protocol compliance or identify important protocol deviations. *Missing values* should be distinguishable from the *value zero* or *characteristic absent*.

The process of data capture, through to database finalization, should be carried out in accordance with good clinical practice (GCP) (see ICH E6, section 5). Specifically, timely and reliable processes for recording data and rectifying errors and omissions are necessary to ensure delivery of a quality database and the achievement of the trial objectives through the implementation of the planned analysis.

IV. TRIAL CONDUCT CONSIDERATIONS

A. Trial Monitoring and Interim Analysis (4.1)

Careful conduct of a clinical trial according to the protocol has a major impact on the credibility of the results (see ICH E6). Careful monitoring can ensure that difficulties are noticed early and their occurrence or recurrence minimized.

There are two distinct types of monitoring that generally characterize confirmatory clinical trials sponsored by the pharmaceutical industry. One type of monitoring concerns the oversight of the quality of the trial, while the other type involves breaking the blind to make treatment comparisons (i.e., interim analysis). Both types of trial monitoring, in addition to entailing different staff responsibilities, involve access to different types of trial data and information, and thus different principles apply for the control of potential statistical and operational bias.

For the purpose of overseeing the quality of the trial, the checks involved in trial monitoring may include whether the protocol is being followed, the acceptability of data being accrued, the success of planned accrual targets, the appropriateness of the design assumptions, success in keeping patients in the trials, and so on (see sections IV.B to IV.D). This type of monitoring does not require access to information on comparative treatment effects nor unblinding of data and, therefore, has no impact on Type I error. The monitoring of a trial for this purpose is the responsibility of the sponsor (see ICH E6) and can be carried out by the sponsor or an independent group selected by the sponsor. The period for this type of monitoring usually starts with the selection of the trial sites and ends with the collection and cleaning of the last subject's data.

The other type of trial monitoring (interim analysis) involves the accruing of comparative treatment results. Interim analysis requires unblinded (i.e., key breaking) access to treatment group assignment (actual treatment assignment or identification of group assignment) and comparative treatment group summary information. Therefore, the protocol (or appropriate amendments prior to a first analysis) should contain statistical plans for the interim analysis to prevent certain types of bias. This is discussed in sections IV.E and IV.F.

B. Changes in Inclusion and Exclusion Criteria (4.2)

Inclusion and exclusion criteria should remain constant, as specified in the protocol, throughout the period of subject recruitment. Changes may occasionally be appropriate, for example, in long-term trials, where growing medical knowledge either from outside the trial or from interim analyses may suggest a change of entry criteria. Changes may also result from the discovery by monitoring staff that regular violations of the entry criteria are occurring or that seriously low recruitment rates are due to over-restrictive criteria. Changes should be made without breaking the blind and should always be described by a protocol amendment. This amendment should cover any statistical consequences, such as sample size adjustments arising from different event rates, or modifications to the planned analysis, such as stratifying the analysis according to modified inclusion/exclusion criteria.

C. Accrual Rates (4.3)

In trials with a long time-scale for the accrual of subjects, the rate of accrual should be monitored. If it falls appreciably below the projected level, the reasons should be identified and remedial actions taken to protect the power of the trial and alleviate concerns about selective entry and other aspects of quality. In a multicenter trial, these considerations apply to the individual centers.

D. Sample Size Adjustment (4.4)

In long-term trials there will usually be an opportunity to check the assumptions which underlie the original design and sample size calculations. This may be particularly important if the trial specifications have been made on preliminary and/or uncertain information. An interim check conducted on the blinded data may reveal that overall response variances, event rates or survival experience are not as anticipated. A revised sample size may then be calculated using suitably modified assumptions, and should be justified and documented in a protocol amendment and in the clinical study report. The steps taken to preserve blindness and the consequences, if any, for the Type I error and the width of confidence intervals should be explained. The potential need for re-estimation of the sample size should be envisaged in the protocol whenever possible (see section III.E).

E. Interim Analysis and Early Stopping (4.5)

An interim analysis is any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to formal completion of a trial. Because the number, methods, and consequences of these comparisons affect the interpretation of the trial, all interim analyses should be carefully planned in advance and described in the protocol. Special circumstances may dictate the need for an interim analysis that was not defined at the start of a trial. In these cases, a protocol amendment describing the interim analysis should be completed prior to unblinded access to treatment comparison data. When an interim analysis is planned with the intention of deciding whether or not to terminate a trial, this is usually accomplished by the use of a group sequential design that employs statistical monitoring schemes as guidelines (see section III.D). The goal of such an interim analysis is to stop the trial early if the superiority of the treatment under study is clearly established, if the demonstration of a relevant treatment difference has become unlikely, or if unacceptable adverse effects are apparent. Generally, boundaries for monitoring efficacy require more evidence to terminate a trial early (i.e., they are more conservative) than boundaries for monitoring safety. When the trial design and monitoring objective involve multiple endpoints, then this aspect of multiplicity may also need to be taken into account.

The protocol should describe the schedule of interim analyses or, at least, the considerations that will govern its generation, for example, if flexible alpha spending function approaches are to be employed. Further details may be given in a protocol amendment before the time of the first interim analysis. The stopping guidelines and their properties should be clearly described in the protocol or amendments. The potential effects of early stopping on the analysis of other important variables should also be considered. This material should be written or approved by the data monitoring committee (see section IV.F), when the trial has one. Deviations from the planned procedure always bear the potential of invalidating the trial results. If it becomes necessary to make changes to the trial, any consequent changes to the statistical procedures should be specified in an amendment to the protocol at the earliest opportunity, especially discussing the impact on any analysis and inferences that such changes may cause. The procedures selected should always ensure that the overall probability of Type I error is controlled.

The execution of an interim analysis should be a completely confidential process because unblinded data and results are potentially involved. All staff involved in the conduct of the trial should remain blind to the results of such analyses, because of the possibility that their attitudes to the trial will be modified and cause changes in the characteristics of patients to be recruited or biases in treatment comparisons. This principle may be applied to all investigator staff and to staff employed by the sponsor except for those who are directly involved in the execution of the interim analysis. Investigators should be informed only about the decision to continue or to discontinue the trial, or to implement modifications to trial procedures.

Most clinical trials intended to support the efficacy and safety of an investigational product should proceed to full completion of planned sample size accrual; trials should be stopped early only for ethical reasons or if the power is no longer acceptable. However, it is recognized that drug development plans involve the need for sponsor access to comparative treatment data for a variety of reasons, such as planning other trials. It is also

recognized that only a subset of trials will involve the study of serious life-threatening outcomes or mortality which may need sequential monitoring of accruing comparative treatment effects for ethical reasons. In either of these situations, plans for interim statistical analysis should be in place in the protocol or in protocol amendments prior to the unblinded access to comparative treatment data in order to deal with the potential statistical and operational bias that may be introduced.

For many clinical trials of investigational products, especially those that have major public health significance, the responsibility for monitoring comparisons of efficacy and/or safety outcomes should be assigned to an external independent group, often called an independent data monitoring committee (IDMC), a data and safety monitoring board, or a data monitoring committee, whose responsibilities should be clearly described.

When a sponsor assumes the role of monitoring efficacy or safety comparisons and therefore has access to unblinded comparative information, particular care should be taken to protect the integrity of the trial and to manage and limit appropriately the sharing of information. The sponsor should ensure and document that the internal monitoring committee has complied with written standard operating procedures and that minutes of decisionmaking meetings, including records of interim results, are maintained.

Any interim analysis that is not planned appropriately (with or without the consequences of stopping the trial early) may flaw the results of a trial and possibly weaken confidence in the conclusions drawn. Therefore, such analyses should be avoided. If unplanned interim analysis is conducted, the clinical study report should explain why it was necessary and the degree to which blindness had to be broken, and provide an assessment of the potential magnitude of bias introduced and the impact on the interpretation of the results.

F. Role of Independent Data Monitoring Committee (IDMC) (4.6)

(see sections 1.25 and 5.5.2 of ICH E6)

An IDMC may be established by the sponsor to assess at intervals the progress of a clinical trial, safety data, and critical efficacy variables and recommend to the sponsor whether to continue, modify or terminate a trial. The IDMC should have written operating procedures and maintain records of all its meetings, including interim results; these should be available for review when the trial is complete. The independence of the IDMC is intended to control the sharing of important comparative information and to protect the integrity of the clinical trial from adverse impact resulting from access to trial information. The IDMC is a separate entity from an institutional review board (IRB) or an independent ethics committee (IEC), and its composition should include clinical trial scientists knowledgeable in the appropriate disciplines, including statistics.

When there are sponsor representatives on the IDMC, their role should be clearly defined in the operating procedures of the committee (e.g., covering whether or not they can vote on key issues). Since these sponsor staff would have access to unblinded information, the

procedures should also address the control of dissemination of interim trial results within the sponsor organization.

V. DATA ANALYSIS CONSIDERATIONS

A. Prespecification of the Analysis (5.1)

When designing a clinical trial, the principal features of the eventual statistical analysis of the data should be described in the statistical section of the protocol. This section should include all the principal features of the proposed confirmatory analysis of the primary variable(s) and the way in which anticipated analysis problems will be handled. In the case of exploratory trials, this section could describe more general principles and directions.

The statistical analysis plan (see Glossary) may be written as a separate document to be completed after finalizing the protocol. In this document, a more technical and detailed elaboration of the principal features stated in the protocol may be included (see section VII.A). The plan may include detailed procedures for executing the statistical analysis of the primary and secondary variables and other data. The plan should be reviewed and possibly updated as a result of the blind review of the data (see section VII.A for definition) and should be finalized before breaking the blind. Formal records should be kept of when the statistical analysis plan was finalized as well as when the blind was subsequently broken.

If the blind review suggests changes to the principal features stated in the protocol, these should be documented in a protocol amendment. Otherwise, it should suffice to update the statistical analysis plan with the considerations suggested from the blind review. Only results from analyses envisaged in the protocol (including amendments) can be regarded as confirmatory.

In the statistical section of the clinical study report, the statistical methodology should be clearly described including when in the clinical trial process methodology decisions were made (see ICH E3).

B. Analysis Sets (5.2)

The set of subjects whose data are to be included in the main analyses should be defined in the statistical section of the protocol. In addition, documentation for all subjects for whom trial procedures (e.g., run-in period) were initiated may be useful. The content of this subject documentation depends on detailed features of the particular trial, but at least demographic and baseline data on disease status should be collected whenever possible.

If all subjects randomized into a clinical trial satisfied all entry criteria, followed all trial procedures perfectly with no losses to follow-up, and provided complete data records, then the set of subjects to be included in the analysis would be self-evident. The design

and conduct of a trial should aim to approach this ideal as closely as possible, but, in practice, it is doubtful if it can ever be fully achieved. Hence**,** the statistical section of the protocol should address anticipated problems prospectively in terms of how these affect the subjects and data to be analyzed. The protocol should also specify procedures aimed at minimizing any anticipated irregularities in study conduct that might impair a satisfactory analysis, including various types of protocol violations, withdrawals and missing values. The protocol should consider ways both to reduce the frequency of such problems and to handle the problems that do occur in the analysis of data. Possible amendments to the way in which the analysis will deal with protocol violations should be identified during the blind review. It is desirable to identify any important protocol violation with respect to the time when it occurred, its cause, and its influence on the trial result. The frequency and type of protocol violations, missing values, and other problems should be documented in the clinical study report and their potential influence on the trial results should be described (see ICH E3).

Decisions concerning the analysis set should be guided by the following principles: (1) To minimize bias and (2) to avoid inflation of Type I error.

1. Full Analysis Set (5.2.1)

The intention-to-treat (see Glossary) principle implies that the primary analysis should include all randomized subjects. Compliance with this principle would necessitate complete follow-up of all randomized subjects for study outcomes. In practice, this ideal may be difficult to achieve, for reasons to be described. In this document, the term *full analysis set* is used to describe the analysis set which is as complete as possible and as close as possible to the intention-to-treat ideal of including all randomized subjects. Preservation of the initial randomization in analysis is important in preventing bias and in providing a secure foundation for statistical tests. In many clinical trials, the use of the full analysis set provides a conservative strategy. Under many circumstances, it may also provide estimates of treatment effects that are more likely to mirror those observed in subsequent practice.

There are a limited number of circumstances that might lead to excluding randomized subjects from the full analysis set, including the failure to satisfy major entry criteria (eligibility violations), the failure to take at least one dose of trial medication, and the lack of any data post randomization. Such exclusions should always be justified. Subjects who fail to satisfy an entry criterion may be excluded from the analysis without the possibility of introducing bias only under the following circumstances:

a. The entry criterion was measured prior to randomization. (i)
b. The detection of the relevant eligibility violations can be made completely objectively. (ii)

c. All subjects receive equal scrutiny for eligibility violations. (This may be difficult to ensure in an open-label study, or even in a double-blind study if the data are unblinded prior to this scrutiny, emphasizing the importance of the blind review.) (iii)

d. All detected violations of the particular entry criterion are excluded. (iv)

In some situations, it may be reasonable to eliminate from the set of all randomized subjects any subject who took no trial medication. The intention-to-treat principle would be preserved despite the exclusion of these patients provided, for example, that the decision of whether or not to begin treatment could not be influenced by knowledge of the assigned treatment. In other situations it may be necessary to eliminate from the set of all randomized subjects any subject without data post randomization. No analysis should be considered complete unless the potential biases arising from these specific exclusions, or any others, are addressed.

When the full analysis set of subjects is used, violations of the protocol that occur after randomization may have an impact on the data and conclusions, particularly if their occurrence is related to treatment assignment. In most respects, it is appropriate to include the data from such subjects in the analysis, consistent with the intention-to-treat principle. Special problems arise in connection with subjects withdrawn from treatment after receiving one or more doses who provide no data after this point, and subjects otherwise lost to follow-up, because failure to include these subjects in the full analysis set may seriously undermine the approach. Measurements of primary variables made at the time of the loss to follow-up of a subject for any reason, or subsequently collected in accordance with the intended schedule of assessments in the protocol, are valuable in this context; subsequent collection is especially important in studies where the primary variable is mortality or serious morbidity. The intention to collect data in this way should be described in the protocol. Imputation techniques, ranging from the carrying forward of the last observation to the use of complex mathematical models, may also be used in an attempt to compensate for missing data. Other methods employed to ensure the availability of measurements of primary variables for every subject in the full analysis set may require some assumptions about the subjects' outcomes or a simpler choice of outcome (e.g., success/failure). The use of any of these strategies should be described and justified in the statistical section of the protocol, and the assumptions underlying any mathematical models employed should be clearly explained. It is also important to demonstrate the robustness of the corresponding results of analysis, especially when the strategy in question could itself lead to biased estimates of treatment effects.

Because of the unpredictability of some problems, it may sometimes be preferable to defer detailed consideration of the manner of dealing with irregularities until the blind review of the data at the end of the trial, and, if so, this should be stated in the protocol.

2. Per Protocol Set (5.2.2)

The *per protocol* set of subjects, sometimes described as the *valid cases*, the *efficac*y sample, or the *evaluable subjects* sample, defines a subset of the subjects in the full analysis set who are more compliant with the protocol and is characterized by criteria such as the following:

a. The completion of a certain prespecified minimal exposure to the treatment regimen (i)
b. The availability of measurements of the primary variable(s) (ii)
c. The absence of any major protocol violations, including the violation of entry criteria (iii)

The precise reasons for excluding subjects from the per protocol set should be fully defined and documented before breaking the blind in a manner appropriate to the circumstances of the specific trial.

The use of the per protocol set may maximize the opportunity for a new treatment to show additional efficacy in the analysis, and most closely reflects the scientific model underlying the protocol. However, the corresponding test of the hypothesis and estimate of the treatment effect may or may not be conservative, depending on the trial. The bias, which may be severe, arises from the fact that adherence to the study protocol may be related to treatment and outcome.

The problems that lead to the exclusion of subjects to create the per protocol set, and other protocol violations, should be fully identified and summarized. Relevant protocol violations may include errors in treatment assignment, the use of excluded medication, poor compliance, loss to followup, and missing data. It is good practice to assess the pattern of such problems among the treatment groups with respect to frequency and time to occurrence.

3. Roles of the Different Analysis Sets (5.2.3)

In general, it is advantageous to demonstrate a lack of sensitivity of the principal trial results to alternative choices of the set of subjects analyzed. In confirmatory trials, it is usually appropriate to plan to conduct both an analysis of the full analysis set and a per protocol analysis, so that any differences between them can be the subject of explicit discussion and interpretation. In some cases, it may be desirable to plan further exploration of the sensitivity of conclusions to the choice of the set of subjects analyzed. When the full analysis set and the per protocol set lead to essentially the same conclusions, confidence in the trial results is increased, bearing in mind, however, that the need to exclude a substantial proportion of subjects from the per protocol analysis throws some doubt on the overall validity of the trial.

The full analysis set and the per protocol set play different roles in superiority trials (which seek to show the investigational product to be superior) and in equivalence or noninferiority trials (which seek to show the investigational product to be comparable, see section III.C.2). In superiority trials, the full analysis set is used in the primary analysis (apart from exceptional circumstances) because it tends to avoid over-optimistic estimates of efficacy resulting from a per protocol analysis. This is because the noncompliers included in the full analysis set will generally diminish the estimated treatment effect. However, in an equivalence or noninferiority trial, use of the full analysis set is generally not conservative and its role should be considered very carefully.

C. Missing Values and Outliers (5.3)

Missing values represent a potential source of bias in a clinical trial. Hence, every effort should be undertaken to fulfill all the requirements of the protocol concerning the collection and management of data. In reality, however, there will almost always be some missing data. A trial may be regarded as valid, nonetheless, provided the methods of dealing with missing values are sensible, particularly if those methods are predefined in the protocol. Definition of methods may be refined by updating this aspect in the statistical analysis plan during the blind review. Unfortunately, no universally applicable methods of handling missing values can be recommended. An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial.

A similar approach should be adopted to exploring the influence of outliers, the statistical definition of which is, to some extent, arbitrary. Clear identification of a particular value as an outlier is most convincing when justified medically as well as statistically, and the medical context will then often define the appropriate action. Any outlier procedure set out in the protocol or the statistical analysis plan should be such as not to favor any treatment group a priori. Once again, this aspect of the analysis can be usefully updated during blind review. If no procedure for dealing with outliers was foreseen in the trial protocol, one analysis with the actual values and at least one other analysis eliminating or reducing the outlier effect should be performed and differences between their results discussed.

D. Data Transformation (5.4)

The decision to transform key variables prior to analysis is best made during the design of the trial on the basis of similar data from earlier clinical trials. Transformations (e.g., square root, logarithm) should be specified in the protocol and a rationale provided, especially for the primary variable(s). The general principles guiding the use of transformations to ensure that the assumptions underlying the statistical methods are met are to be found in standard texts; conventions for particular variables have been developed in a number of specific clinical areas. The decision on whether and how to transform a variable should be influenced by the preference for a scale that facilitates clinical interpretation.

Similar considerations apply to other derived variables, such as the use of change from baseline, percentage change from baseline, the *area under the curve* of repeated measures, or the ratio of two different variables. Subsequent clinical interpretation should be carefully considered, and the derivation should be justified in the protocol. Closely related points are made in section II.B.2.

E. Estimation, Confidence Intervals, and Hypothesis Testing (5.5)

The statistical section of the protocol should specify the hypotheses that are to be tested and/or the treatment effects that are to be estimated in order to satisfy the primary objectives of the trial. The statistical methods to be used to accomplish these tasks should be described for the primary (and preferably the secondary) variables, and the underlying statistical model should be made clear. Estimates of treatment effects should be accompanied by confidence intervals, whenever possible, and the way in which these will be calculated should be identified. A description should be given of any intentions to use baseline data to improve precision or to adjust estimates for potential baseline differences, for example, by means of analysis of covariance.

It is important to clarify whether one- or two-sided tests of statistical significance will be used and, in particular, to justify prospectively the use of one-sided tests. If hypothesis tests are not considered appropriate, then the alternative process for arriving at statistical conclusions should be given. The issue of one-sided or two-sided approaches to inference is controversial, and a diversity of views can be found in the statistical literature. The approach of setting Type I errors for one-sided tests at half the conventional Type I error used in two-sided tests is preferable in regulatory settings. This promotes consistency with the two-sided confidence intervals that are generally appropriate for estimating the possible size of the difference between two treatments.

The particular statistical model chosen should reflect the current state of medical and statistical knowledge about the variables to be analyzed as well as the statistical design of the trial. All effects to be fitted in the analysis (for example, in analysis of variance models) should be fully specified, and the manner, if any, in which this set of effects might be modified in response to preliminary results should be explained. The same considerations apply to the set of covariates fitted in an analysis of covariance. (See also section V.G.) In the choice of statistical methods, due attention should be paid to the statistical distribution of both primary and secondary variables. When making this choice (for example between parametric and nonparametric methods), it is important to bear in mind the need to provide statistical estimates of the size of treatment effects together with confidence intervals (in addition to significance tests).

The primary analysis of the primary variable should be clearly distinguished from supporting analyses of the primary or secondary variables. Within the statistical section of the protocol or the statistical analysis plan there should also be an outline of the way in which data other than the primary and secondary variables will be summarized and

reported. This should include a reference to any approaches adopted for the purpose of achieving consistency of analysis across a range of trials, for example, for safety data.

Modeling approaches that incorporate information on known pharmacological parameters, the extent of protocol compliance for individual subjects, or other biologically based data may provide valuable insights into actual or potential efficacy, especially with regard to estimation of treatment effects. The assumptions underlying such models should always be clearly identified, and the limitations of any conclusions should be carefully described.

F. Adjustment of Significance and Confidence Levels (5.6)

When multiplicity is present, the usual frequentist approach to the analysis of clinical trial data may necessitate an adjustment to the Type I error. Multiplicity may arise, for example, from multiple primary variables (see section II.B.2), multiple comparisons of treatments, repeated evaluation over time, and/or interim analyses (see section IV.E). Methods to avoid or reduce multiplicity are sometimes preferable when available, such as the identification of the key primary variable (multiple variables), the choice of a critical treatment contrast (multiple comparisons), and the use of a summary measure such as *area under the curve* (repeated measures). In confirmatory analyses, any aspects of multiplicity that remain after steps of this kind have been taken should be identified in the protocol; adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.

G. Subgroups, Interactions, and Covariates (5.7)

The primary variable(s) is often systematically related to other influences apart from treatment. For example, there may be relationships to covariates such as age and sex, or there may be differences between specific subgroups of subjects, such as those treated at the different centers of a multicenter trial. In some instances, an adjustment for the influence of covariates or for subgroup effects is an integral part of the planned analysis and hence should be set out in the protocol. Pretrial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis to improve precision and to compensate for any lack of balance between treatment groups. If one or more factors are used to stratify the design, it is appropriate to account for those factors in the analysis. When the potential value of an adjustment is in doubt, it is often advisable to nominate the unadjusted analysis as the one for primary attention, the adjusted analysis being supportive. Special attention should be paid to center effects and to the role of baseline measurements of the primary variable. It is not advisable to adjust the main analyses for covariates measured after randomization because they may be affected by the treatments.

The treatment effect itself may also vary with subgroup or covariate. For example, the effect may decrease with age or may be larger in a particular diagnostic category of

subjects. In some cases such interactions are anticipated or are of particular prior interest (e.g., geriatrics); hence a subgroup analysis or a statistical model including interactions is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall. In general, such analyses should proceed first through the addition of interaction terms to the statistical model in question, complemented by additional exploratory analysis within relevant subgroups of subjects, or within strata defined by the covariates. When exploratory, these analyses should be interpreted cautiously. Any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses is unlikely to be accepted.

H. Integrity of Data and Computer Software Validity (5.8)

The credibility of the numerical results of the analysis depends on the quality and validity of the methods and software (both internally and externally written) used both for data management (data entry, storage, verification, correction, and retrieval) and for processing the data statistically. Data management activities should therefore be based on thorough and effective standard operating procedures. The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.

VI. EVALUATION OF SAFETY AND TOLERABILITY

A. Scope of Evaluation (6.1)

In all clinical trials, evaluation of safety and tolerability (see Glossary) constitutes an important element. In early phases this evaluation is mostly of an exploratory nature and is only sensitive to frank expressions of toxicity, whereas in later phases the establishment of the safety and tolerability profile of a drug can be characterized more fully in larger samples of subjects. Later phase controlled trials represent an important means of exploring, in an unbiased manner, any new potential adverse effects, even if such trials generally lack power in this respect.

Certain trials may be designed with the purpose of making specific claims about superiority or equivalence with regard to safety and tolerability compared to another drug or to another dose of the investigational drug. Such specific claims should be supported by relevant evidence from confirmatory trials, similar to that necessary for corresponding efficacy claims.

B. Choice of Variables and Data Collection (6.2)

In any clinical trial, the methods and measurements chosen to evaluate the safety and tolerability of a drug will depend on a number of factors, including knowledge of the adverse effects of closely related drugs, information from nonclinical and earlier clinical trials and possible consequences of the pharmacodynamic/ pharmacokinetic properties of

the particular drug, the mode of administration, the type of subjects to be studied, and the duration of the trial. Laboratory tests concerning clinical chemistry and hematology, vital signs, and clinical adverse events (diseases, signs, and symptoms) usually form the main body of the safety and tolerability data. The occurrence of serious adverse events and treatment discontinuations due to adverse events are particularly important to register (see ICH E2A and ICH E3).

Furthermore, it is recommended that a consistent methodology be used for the data collection and evaluation throughout a clinical trial program to facilitate the combining of data from different trials. The use of a common adverse event dictionary is particularly important. This dictionary has a structure that makes it possible to summarize the adverse event data on three different levels: System-organ class, preferred term, or included term (see Glossary). The preferred term is the level on which adverse events usually are summarized, and preferred terms belonging to the same system-organ class could then be brought together in the descriptive presentation of data (see ICH M1).

C. Set of Subjects to Be Evaluated and Presentation of Data (6.3)

For the overall safety and tolerability assessment, the set of subjects to be summarized is usually defined as those subjects who received at least one dose of the investigational drug. Safety and tolerability variables should be collected as comprehensively as possible from these subjects, including type of adverse event, severity, onset, and duration (see ICH E2B). Additional safety and tolerability evaluations may be needed in specific subpopulations, such as females, the elderly (see ICH E7), the severely ill, or those who have a common concomitant treatment. These evaluations may need to address more specific issues (see ICH E3).

All safety and tolerability variables will need attention during evaluation, and the broad approach should be indicated in the protocol. All adverse events should be reported, whether or not they are considered to be related to treatment. All available data in the study population should be accounted for in the evaluation. Definitions of measurement units and reference ranges of laboratory variables should be made with care; if different units or different reference ranges appear in the same trial (e.g., if more than one laboratory is involved), then measurements should be appropriately standardized to allow a unified evaluation. Use of a toxicity grading scale should be prespecified and justified.

The incidence of a certain adverse event is usually expressed in the form of a proportion relating number of subjects experiencing events to number of subjects at risk. However, it is not always self-evident how to assess incidence. For example, depending on the situation, the number of exposed subjects or the extent of exposure (in person-years) could be considered for the denominator. Whether the purpose of the calculation is to estimate a risk or to make a comparison between treatment groups, it is important that the definition is given in the protocol. This is especially important if long-term treatment is planned and a substantial proportion of treatment withdrawals or deaths are expected. For such

situations, survival analysis methods should be considered and cumulative adverse event rates calculated in order to avoid the risk of underestimation.

In situations when there is a substantial background noise of signs and symptoms (e.g., in psychiatric trials), one should consider ways for accounting for this in the estimation of risk for different adverse events. One such method is to make use of the *treatment emergent* (see Glossary) concept in which adverse events are recorded only if they emerge or worsen relative to pretreatment baseline.

Other methods to reduce the effect of the background noise may also be appropriate, such as ignoring adverse events of mild severity or requiring that an event should have been observed at repeated visits to qualify for inclusion in the numerator. Such methods should be explained and justified in the protocol.

D. Statistical Evaluation (6.4)

The investigation of safety and tolerability is a multidimensional problem. Although some specific adverse effects can usually be anticipated and specifically monitored for any drug, the range of possible adverse effects is very large, and new and unforeseeable effects are always possible. Further, an adverse event experienced after a protocol violation, such as use of an excluded medication, may introduce a bias. This background underlies the statistical difficulties associated with the analytical evaluation of safety and tolerability of drugs, and means that conclusive information from confirmatory clinical trials is the exception rather than the rule.

In most trials, the safety and tolerability implications are best addressed by applying descriptive statistical methods to the data, supplemented by calculation of confidence intervals wherever this aids interpretation. It is also valuable to make use of graphical presentations in which patterns of adverse events are displayed both within treatment groups and within subjects.

The calculation of p-values is sometimes useful, either as an aid to evaluating a specific difference of interest or as a *flagging* device applied to a large number of safety and tolerability variables to highlight differences worthy of further attention. This is particularly useful for laboratory data, which otherwise can be difficult to summarize appropriately. It is recommended that laboratory data be subjected to both a quantitative analysis (e.g., evaluation of treatment means) and a qualitative analysis where counting of numbers above or below certain thresholds are calculated.

If hypothesis tests are used, statistical adjustments for multiplicity to quantify the Type I error are appropriate, but the Type II error is usually of more concern. Care should be taken when interpreting putative statistically significant findings when there is no multiplicity adjustment.

In the majority of trials, investigators are seeking to establish that there are no clinically unacceptable differences in safety and tolerability compared with either a comparator drug or a placebo. As is the case for noninferiority or equivalence evaluation of efficacy, the use of confidence intervals is preferred to hypothesis testing in this situation. In this way, the considerable imprecision often arising from low frequencies of occurrence is clearly demonstrated.

E. Integrated Summary (6.5)

The safety and tolerability properties of a drug are commonly summarized across trials continuously during an investigational product's development and, in particular, at the time of a marketing application. The usefulness of this summary, however, is dependent on adequate and well-controlled individual trials with high data quality.

The overall usefulness of a drug is always a question of balance between risk and benefit. In a single trial, such a perspective could also be considered even if the assessment of risk/benefit usually is performed in the summary of the entire clinical trial program (see section VII.B.2).

For more details on the reporting of safety and tolerability, see section 12 of ICH E3.

VII. REPORTING

A. Evaluation and Reporting (7.1)

As stated in the introduction, the structure and content of clinical study reports is the subject of ICH E3. That ICH guidance fully covers the reporting of statistical work, appropriately integrated with clinical and other material. The current section is therefore relatively brief.

During the planning phase of a trial, the principal features of the analysis should have been specified in the protocol as described in section V. When the conduct of the trial is over and the data are assembled and available for preliminary inspection, it is valuable to carry out the blind review of the planned analysis also described in section V. This pre-analysis review, blinded to treatment, should cover, for example, decisions concerning the exclusion of subjects or data from the analysis sets, the checking of possible transformations and definitions of outliers, the addition to the model of important covariates identified in other recent research, and the reconsideration of the use of parametric or nonparametric methods. Decisions made at this time should be described in the report and should be distinguished from those made after the statistician has had access to the treatment codes, as blind decisions will generally introduce less potential for bias. Statisticians or other staff involved in unblinded interim analysis should not participate in the blind review or in making modifications to the statistical analysis plan. When the blinding is compromised by the possibility that treatment-induced effects may be apparent in the data, special care will be needed for the blind review.

Many of the more detailed aspects of presentation and tabulation should be finalized at or about the time of the blind review so that, by the time of the actual analysis, full plans exist for all its aspects including subject selection, data selection and modification, data summary and tabulation, estimation, and hypothesis testing. Once data validation is complete, the analysis should proceed according to the predefined plans; the more these plans are adhered to, the greater the credibility of the results. Particular attention should be paid to any differences between the planned analysis and the actual analysis as described in the protocol, the protocol amendments, or the updated statistical analysis plan based on a blind review of data. A careful explanation should be provided for deviations from the planned analysis.

All subjects who entered the trial should be accounted for in the report, whether or not they are included in the analysis. All reasons for exclusion from analysis should be documented; for any subject included in the full analysis set but not in the per protocol set, the reasons for exclusion from the latter should also be documented. Similarly, for all subjects included in an analysis set, the measurements of all important variables should be accounted for at all relevant time-points.

The effect of all losses of subjects or data, withdrawals from treatment, and major protocol violations on the main analyses of the primary variable(s) should be considered carefully. Subjects lost to followup, withdrawn from treatment, or with a severe protocol violation should be identified and a descriptive analysis of them provided, including the reasons for their loss and its relationship to treatment and outcome.

Descriptive statistics form an indispensable part of reports. Suitable tables and/or graphical presentations should illustrate clearly the important features of the primary and secondary variables and of key prognostic and demographic variables. The results of the main analyses relating to the objectives of the trial should be the subject of particularly careful descriptive presentation. When reporting the results of significance tests, precise p-values (e.g., $p=0.034$) should be reported rather than making exclusive reference to critical values.

Although the primary goal of the analysis of a clinical trial should be to answer the questions posed by its main objectives, new questions based on the observed data may well emerge during the unblinded analysis. Additional and perhaps complex statistical analysis may be the consequence. This additional work should be strictly distinguished in the report from work which was planned in the protocol.

The play of chance may lead to unforeseen imbalances between the treatment groups in terms of baseline measurements not predefined as covariates in the planned analysis but having some prognostic importance nevertheless. This is best dealt with by showing that an additional analysis which accounts for these imbalances reaches essentially the same conclusions as the planned analysis. If this is not the case, the effect of the imbalances on the conclusions should be discussed.

In general, sparing use should be made of unplanned analyses. Such analyses are often carried out when it is thought that the treatment effect may vary according to some other factor or factors. An attempt may then be made to identify subgroups of subjects for whom the effect is particularly beneficial. The potential dangers of over-interpretation of unplanned subgroup analyses are well known (see also section V.G) and should be carefully avoided. Although similar problems of interpretation arise if a treatment appears to have no benefit or an adverse effect in a subgroup of subjects, such possibilities should be properly assessed and should therefore be reported.

Finally, statistical judgement should be brought to bear on the analysis, interpretation and presentation of the results of a clinical trial. To this end, the trial statistician should be a member of the team responsible for the clinical study report and should approve the clinical report.

B. Summarizing the Clinical Database (7.2)

An overall summary and synthesis of the evidence on safety and efficacy from all the reported clinical trials is required for a marketing application (*expert report* in EU, *integrated summary* reports in the United States, *gaiyou* in Japan). This may be accompanied, when appropriate, by a statistical combination of results.

Within the summary a number of areas of specific statistical interest arise: Describing the demography and clinical features of the population treated during the course of the clinical trial program; addressing the key questions of efficacy by considering the results of the relevant (usually controlled) trials and highlighting the degree to which they reinforce or contradict each other; summarizing the safety information available from the combined database of all the trials whose results contribute to the marketing application; and identifying potential safety issues. During the design of a clinical program, careful attention should be paid to the uniform definition and collection of measurements which will facilitate subsequent interpretation of the series of trials, particularly if they are likely to be combined across trials. A common dictionary for recording the details of medication, medical history and adverse events should be selected and used. A common definition of the primary and secondary variables is nearly always worthwhile and is essential for meta-analysis. The manner of measuring key efficacy variables, the timing of assessments relative to randomization/entry, the handling of protocol violators and deviators, and perhaps the definition of prognostic factors should all be kept compatible unless there are valid reasons not to do so.

Any statistical procedures used to combine data across trials should be described in detail. Attention should be paid to the possibility of bias associated with the selection of trials, to the homogeneity of their results, and to the proper modelling of the various sources of variation. The sensitivity of conclusions to the assumptions and selections made should be explored.

1. Efficacy Data (7.2.1)

Individual clinical trials should always be large enough to satisfy their objectives. Additional valuable information may also be gained by summarizing a series of clinical trials that address essentially identical key efficacy questions. The main results of such a set of trials should be presented in an identical form to permit comparison, usually in tables or graphs that focus on estimates plus confidence limits. The use of meta-analytic techniques to combine these estimates is often a useful addition because it allows a more precise overall estimate of the size of the treatment effects to be generated and provides a complete and concise summary of the results of the trials. Under exceptional circumstances, a meta-analytic approach may also be the most appropriate way, or the only way, of providing sufficient overall evidence of efficacy via an overall hypothesis test. When used for this purpose, the meta-analysis should have its own prospectively written protocol.

2. Safety Data (7.2.2)

In summarizing safety data, it is important to examine the safety database thoroughly for any indications of potential toxicity and to follow up any indications by looking for an associated supportive pattern of observations. The combination of the safety data from all human exposure to the drug provides an important source of information because its larger sample size provides the best chance of detecting the rarer adverse events and, perhaps, of estimating their approximate incidence. However, incidence data from this database are difficult to evaluate because of the lack of a comparator group, and data from comparative trials are especially valuable in overcoming this difficulty. The results from trials which use a common comparator (placebo or specific active comparator) should be combined and presented separately for each comparator providing sufficient data.

All indications of potential toxicity arising from exploration of the data should be reported. The evaluation of the reality of these potential adverse effects should take into account the issue of multiplicity arising from the numerous comparisons made. The evaluation should also make appropriate use of survival analysis methods to exploit the potential relationship of the incidence of adverse events to duration of exposure and/or followup. The risks associated with identified adverse effects should be appropriately quantified to allow a proper assessment of the risk/benefit relationship.

GLOSSARY (Annex 1)

Bayesian approaches: Approaches to data analysis that provide a posterior probability distribution for some parameter (e.g., treatment effect), derived from the observed data and a prior probability distribution for the parameter. The posterior distribution is then used as the basis for statistical inference.

Bias (statistical and operational): The systematic tendency of any factors associated with the design, conduct, analysis and evaluation of the results of a clinical trial to make the estimate of a treatment effect deviate from its true value. Bias introduced through deviations in conduct is referred to as *operational bias*. The other sources of bias listed above are referred to as *statistical bias*.

Blind review: The checking and assessment of data during the period of time between trial completion (the last observation on the last subject) and the breaking of the blind, for the purpose of finalizing the planned analysis.

Content validity: The extent to which a variable (e.g., a rating scale) measures what it is supposed to measure.

Double dummy: A technique for retaining the blind when administering supplies in a clinical trial, when the two treatments cannot be made identical. Supplies are prepared for Treatment A (active and indistinguishable placebo) and for Treatment B (active and indistinguishable placebo). Subjects then take two sets of treatment; either A (active) and B (placebo), or A (placebo) and B (active).

Dropout: A subject in a clinical trial who for any reason fails to continue in the trial until the last visit required of him/her by the study protocol.

Equivalence trial: A trial with the primary objective of showing that the response to two or more treatments differs by an amount which is clinically unimportant. This is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence margin of clinically acceptable differences.

Frequentist methods: Statistical methods, such as significance tests and confidence intervals, which can be interpreted in terms of the frequency of certain outcomes occurring in hypothetical repeated realizations of the same experimental situation.

Full analysis set: The set of subjects that is as close as possible to the ideal implied by the intention-to-treat principle. It is derived from the set of all randomized subjects by minimal and justified elimination of subjects.

Generalizability, generalization: The extent to which the findings of a clinical trial can be reliably extrapolated from the subjects who participated in the trial to a broader patient population and a broader range of clinical settings.

Global assessment variable: A single variable, usually a scale of ordered categorical ratings, that integrates objective variables and the investigator's overall impression about the state or change in state of a subject.

Independent data monitoring committee (IDMC) (data and safety monitoring board, monitoring committee, data monitoring committee): An independent data monitoring committee that may be established by the sponsor to assess at intervals the progress of a clinical trial, the safety data, and the critical efficacy endpoints, and to recommend to the sponsor whether to continue, modify, or stop a trial.

Intention-to-treat principle: The principle that asserts that the effect of a treatment policy can be best assessed by evaluating on the basis of the intention to treat a subject (i.e., the planned treatment regimen) rather than the actual treatment given. It has the consequence that subjects allocated to a treatment group should be followed up, assessed, and analyzed as members of that group irrespective of their compliance with the planned course of treatment.

Interaction (qualitative and quantitative): The situation in which a treatment contrast (e.g., difference between investigational product and control) is dependent on another factor (e.g., center). A quantitative interaction refers to the case where the magnitude of the contrast differs at the different levels of the factor, whereas for a qualitative interaction the direction of the contrast differs for at least one level of the factor.

Interrater reliability: The property of yielding equivalent results when used by different raters on different occasions.

Intrarater reliability: The property of yielding equivalent results when used by the same rater on different occasions.

Interim analysis: Any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to the formal completion of a trial.

Meta-analysis: The formal evaluation of the quantitative evidence from two or more trials bearing on the same question. This most commonly involves the statistical combination of summary statistics from the various trials, but the term is sometimes also used to refer to the combination of the raw data.

Multicenter trial: A clinical trial conducted according to a single protocol but at more than one site and, therefore, carried out by more than one investigator.

Noninferiority trial: A trial with the primary objective of showing that the response to the investigational product is not clinically inferior to a comparative agent (active or placebo control).

Preferred and included terms: In a hierarchical medical dictionary, for example, the World Health Organization's Adverse Reaction Terminology (WHO-Art), the included term is the lowest level of dictionary term to which the investigator description is coded. The preferred term is the level of grouping of included terms typically used in reporting frequency of occurrence. For example, the investigator text "Pain in the left arm" might be coded to the included term "Joint pain," which is reported at the preferred term level as "Arthralgia."

Per protocol set (valid cases, efficacy sample, evaluable subjects sample): The set of data generated by the subset of subjects who complied with the protocol sufficiently to ensure that these data would be likely to exhibit the effects of treatment according to the underlying scientific model. Compliance covers such considerations as exposure to treatment, availability of measurements, and absence of major protocol violations.

Safety and tolerability: The safety of a medical product concerns the medical risk to the subject, usually assessed in a clinical trial by laboratory tests (including clinical chemistry and hematology), vital signs, clinical adverse events (diseases, signs and symptoms), and other special safety tests (e.g., electrocardiograms, ophthalmology). The tolerability of the medical product represents the degree to which overt adverse effects can be tolerated by the subject.

Statistical analysis plan: A statistical analysis plan is a document that contains a more technical and detailed elaboration of the principal features of the analysis described in the protocol, and includes detailed procedures for executing the statistical analysis of the primary and secondary variables and other data.

Superiority trial: A trial with the primary objective of showing that the response to the investigational product is superior to a comparative agent (active or placebo control).

Surrogate variable: A variable that provides an indirect measurement of effect in situations where direct measurement of clinical effect is not feasible or practical.

Treatment effect: An effect attributed to a treatment in a clinical trial. In most clinical trials, the treatment effect of interest is a comparison (or contrast) of two or more treatments.

Treatment emergent: An event that emerges during treatment, having been absent pretreatment, or worsens relative to the pretreatment state.

Trial statistician: A statistician who has a combination of education/training and experience sufficient to implement the principles in this guidance and who is responsible for the statistical aspects of the trial.