



London, 17 March 2005
CHMP/EWP/83561/2005

**COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
(CHMP)**

DRAFT

GUIDELINE ON CLINICAL TRIALS IN SMALL POPULATIONS

DISCUSSION IN THE EFFICACY WORKING PARTY / AD HOC GROUP ON CLINICAL TRIALS IN SMALL POPULATIONS	May 2002 – January 2005
TRANSMISSION TO CHMP	March 2005
RELEASE FOR CONSULTATION	March 2005
DEADLINE FOR COMMENTS	September 2005

Note:

Any comments to this Guideline should be sent to the EMEA EWP Secretariat by e-mail: juan.garcia@emea.eu.int or by fax: +44 20 74 18 86 13 by the end of September 2005

GUIDELINE ON CLINICAL TRIALS IN SMALL POPULATIONS

TABLE OF CONTENTS

1. Introduction	3
2. Levels of Evidence	4
3. Pharmacological Considerations	5
4. Choice of Endpoints	6
5. Choice of Control Groups	8
6. Methodological and Statistical Considerations	8
7. Summary and Conclusions	13
Appendix	15

1. Introduction

This Discussion Paper considers problems associated with clinical trials when there are very few patients available to study. Many rare diseases affect only a few thousand or even fewer than one hundred patients in the EU. Under such circumstances a trial enrolling several hundred patients may not be practical or possible. Accordingly, conduct, analysis, and interpretation of studies in rare conditions may at times be constrained to varying degrees by the prevalence of the disease.

The paper has been prepared by the CHMP Efficacy Working Party (EWP) in joint collaboration with members of the Scientific Advice Working Party (SAWP), the Committee on Orphan Medicinal Products (COMP) and the Paediatric Expert Group (PEG). The expertise within the group includes clinicians, epidemiologists and statisticians from National Regulatory Authorities and from universities.

No methods exist that are relevant to small studies that are not also applicable to large studies. However, it may be that in conditions with small and very small populations, less conventional and/or less commonly seen methodological approaches may be acceptable. In this document strategies for an approach to trials are briefly outlined.

Some of the approaches outlined in this document are primarily intended for situations where large studies are not feasible. They should not be interpreted as a general paradigm change in the evaluation of drug development.

Decisions taken during the process of marketing authorisation of medicinal products are always uncertain. Evidence that is 'beyond doubt' never exists. Patients, scientists, regulators and pharmaceutical companies must accept that it is hardly ever possible to *prove* any claim, particularly not with the largely inductive approach taken by the medical sciences. Regulators therefore prefer well-used and reliable methods.

Because of the shortfalls of using inductive inference (that is, drawing general conclusions from specific cases), great care must be taken to follow established guidelines for the conduct of clinical trials whenever possible. This present guideline is exclusively intended for situations where such established guidelines cannot be followed.

In general, methods to increase the efficiency of the design and analysis are also applicable for studies in large populations but are not often used because of increased complexity. As mentioned above, the general principle can be applied to two types of situations: (1) when randomised controlled trials are feasible, even though the interpretation will be less clear compared to typical phase III trials containing several hundreds or even thousands of patients, (2) when randomised controlled trials will be severely underpowered if feasible at all. In this situation, case series (with external control groups) and sometimes only anecdotal cases reports are available.

These two situations may arise in the field of rare diseases but also in rapidly evolving fields (like organ transplantation). Further, refinement of individually targeted medicinal products, e.g. by applying pharmacogenomics, may lead to many more, but smaller, target populations. The examples given in this paper may not be considered appropriate in some conventional situations but may be applicable to others. In situations where only very few patients can be enrolled in studies, alternative approaches are required. Such compromise positions will usually be at the cost of increased uncertainty concerning the reliability of the results and hence the reliability of the effectiveness, safety and risk–benefit of the product. However several orphan products have been authorised even though randomised controlled trials have not been performed.

This document addresses (1) methods where the efficiency of the design or analysis may be increased, and (2) approaches for situations where such methods are not applicable. General principles and specific examples are presented. Such outlines and examples are by no means exhaustive but should encourage further exploration of potentially suitable methods for specific situations. It should be read in conjunction with the following Directives and documents:

- Directive 2001/20/EC of the European Parliament and Council of 4 April 2001 on the approximation of the Laws, regulations and administrative provisions of the Member States

relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use.

- Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use.
- The Extent of Population Exposure to Assess Clinical Safety for Drugs (ICH E1A).
- Dose Response Information to Support Drug Registration (ICH E4).
- General Considerations for Clinical Trials (ICH E8).
- Statistical Principles for Clinical Trials (ICH E9).
- Choice of Control Group in Clinical Trials (ICH E10).
- Clinical Investigation of Medicinal Products in the Paediatric Population (ICH E11).
- Accelerated Evaluation of Products Indicated for Serious Diseases (Life Threatening or Heavily Disabling Diseases) (CPMP/495/96 rev. 1).
- Points to Consider on the Evaluation of the Diagnostic Agents (CPMP/EWP/1119/98).
- Points to Consider on Applications with 1.) Meta-analyses and 2.) One Pivotal Study (CPMP/2330/99).
- Points to Consider on Calculation and Reporting of the Prevalence of a Condition for Orphan Designation (CPMP/436/01).
- Note for Sponsors of Orphan Medicinal Products Regarding Enlargement of the European Union (EMEA/35607/03).

2. Levels of Evidence

Applications for marketing authorisations in small populations will be judged against the same standards as for other products, although limitations on patient recruitment will be taken into account.

Hierarchies of evidence have been described which usually place in order:

- Meta-analyses of randomised controlled clinical trials
- Individual randomised controlled trials
- Meta-analyses of observational studies
- Individual observational studies
- Published case-reports
- Anecdotal case-reports
- Opinion of experts in the field.

All such forms of evidence provide some information (even anecdotal case reports) and none should be ignored. However, high levels of evidence in drug development come from well-planned and well-executed controlled clinical trials, particularly trials that have minimised bias through appropriate blinding and randomisation. At their conclusion, the treatment effect should ideally be large, confidence intervals for that effect narrow, and the effect size highly statistically significant. Well-planned and well-conducted meta-analyses of such trials provide even stronger evidence. It must be recognised, however, that poor meta-analyses will not give reliable conclusions.

In very rare diseases, the combination of single case studies may be the only way to accumulate evidence. In such situations, treatment regimens and data collection may still be carried out in a controlled manner and this will add weight to the evidence. Furthermore, if careful consideration is given to the statistical analysis, (including methods such as formal 'cumulative meta-analyses' of randomised controlled trials) then this will carry more weight than ad hoc pooling of several case-reports. Meta-analyses of individual case reports or of observational studies should be considered.

Generally, a larger sample size and/or a smaller variance will result in narrower confidence intervals and more extreme levels of statistical significance. The chance of producing a ‘statistically significant’ result (whether the treatment is effective, *or not*) is increased by using a less extreme significance level (for example, considering $P < 0.10$ rather than $P < 0.05$ as the threshold for ‘statistical significance’). Note, also, the 0.05 is a common but arbitrary threshold. No such value is adequate to confirm that a treatment effect truly does exist. Different significance levels may be acceptable on a case-by-case basis but should always be prospectively justified. In almost all cases, confidence intervals of estimates of the treatment effect are much more informative than P -values.

3. Pharmacological Considerations

Detailed knowledge of the pathophysiology of the disease and the pharmacology of the drug will facilitate the design of efficient clinical studies and will help determine the amount of clinical data required.

Pre-clinical pharmacology studies are of special importance for rare diseases and can frequently be used to inform the design of clinical studies. Such studies may also give important information regarding features such as dosing, dose frequency, route of administration, and so on, although investigation of these in man is still preferable whenever possible.

For ‘substitution studies’ (typically enzyme or hormone replacement), well-characterised short- and long-term consequences of the deficiency, and a clear understanding of the pharmacokinetics and pharmacodynamics of the compound, provide guidance for designing studies. Regulatory requirements for licensing ‘substitution products’ may sometimes be less rigid than for other compounds provided that symptoms related to the deficiency are clearly understood and that the pharmacokinetics and pharmacodynamics of the product are well documented in clinical studies. Within-patient comparisons in a relentlessly and predictably progressive disorder might provide sufficient data to support a benefit–risk assessment.

Variability (whether in terms of disease phenotype, underlying pathophysiology, pharmacology or pharmacokinetics) is a threat to successful drug development. Efficient study design and analysis requires as clear an understanding as possible of all of these potential sources of variability.

The credibility of study results may be enhanced if a clear dose-response relationship is seen or in cases where a clear chain of events can be identified (for example, drug exposure to target occupancy, to dynamic measures, to clinical outcome). ‘Black box designs’, on the other hand, are much less convincing and will increase the data requirements regarding robustness and persuasiveness of study results.

In very rare disorders, it is important that every patient participating in a study contributes as much information as possible to help make a benefit–risk assessment possible. Therefore, the well-planned use of the best available techniques to obtain and analyse information is crucial. This applies throughout the study process from pharmacokinetic and pharmacodynamic modelling to handling and analyses of biopsy material.

4. Choice of Endpoints

Ideally a ‘hard’ and clinically relevant endpoint is used. At one extreme, the endpoint may be complete ‘cure’ of disease.

Slowing disease progression is an intermediate level of endpoint and it must be possible to define a measure of disease severity or of disease progression. Ideally, this should be validated as a tool for use in clinical trials but it is recognised that there might be too few patients to use some for validating endpoints and others for testing treatments. In studies whose endpoint is time to progression or time to remission, adequate long-term followed up of patients is important; such evidence will often be from ‘open-label extensions’ to randomised studies. It is preferable, for example, to be able to identify whether a treatment does *cause* a particular (beneficial) outcome, or whether it just delays disease progression.

Clinical endpoints like renal failure (e.g. in Fabry's disease), is a good example of a clinically highly relevant endpoint because it may severely impair a patient's survival and well being. *Relief of symptoms* is also a useful clinical endpoint – usually highly recognised by patients – but it may not reflect slowing true disease progression or delaying death. Even in the absence of demonstration of benefit on a clinical endpoint, relief of symptoms and the resulting patient preference may be a valuable study endpoint. However, this must always be on a disease and treatment-dependent justification.

Improving clinical endpoints may not necessarily be sufficient when patients remain severely disabled (such as poor neurological status following resuscitation or after an intracranial bleed). If *quality of life* is measured, it should always be assessed using scales validated for the particular indication being treated although, as commented earlier, it is recognised that there may sometimes be too few patients for validation exercises as well as separate treatment evaluation. Even with this restriction, improvements in quality of life alone (that is, in the absence of any other clinical benefit) might not be sufficient to grant a Marketing Authorisation. Quality of life data should ideally be considered as supportive evidence. It may be one means of helping to place the product in context with other available treatments.

In some cases, the choice of 'most appropriate' clinical endpoint may not be widely agreed. In other cases, the mode of action of the test treatment may not be well enough known to predict which of several possible outcomes will be affected. In such circumstances, the usual approach of pre-specifying the primary endpoint may be too conservative and more knowledge may be gained from collecting all sensible/possible endpoints and then presenting all the data in the final study report. Still, every effort should be made to identify an appropriate hierarchy in the endpoints. If, collectively, the data look compelling, then a Marketing Authorisation may be grantable.

If recruitment of a sufficient number of patients seems jeopardized or if it would take an unreasonable length of time, then a surrogate marker may offer a solution. A biomarker is a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures how a patient feels, functions, or survives. Preferably the term 'surrogate endpoint' should only be used for biomarkers, which have been validated. In the context of rare disorders, considering a biomarker as a valid surrogate endpoint requires it to be reasonably likely – based on epidemiologic, pathophysiologic, or other evidence – to predict benefit. For example, the number of blasts or BCR/ABL gene products serves as good prognostic indicators for chronic myeloid leukaemia. Thus haematological and cytogenetic response to treatment is considered a valid surrogate endpoint. Other examples include CD4 cell counts and HIV viral load as surrogates for death or opportunistic infections in the evaluation of antiviral agents. Prediction in itself may not, however, be sufficient to attain the status of surrogate and a surrogate marker may not be sufficient to establish efficacy. Considerations should include:

- How closely changes in the surrogate endpoint are linked to causing changes in a clinical or symptomatic endpoint
- How much risk is associated with the therapy
- What other therapies (if any) are available for the same condition.

Validation of surrogate endpoints is difficult. Epidemiological data and data from patient registers may provide sources for validation of potential surrogate markers of disease. This may be of limited value when there are very few patients. Surrogate endpoints will always have the disadvantage of being difficult to relate to real clinical benefit and when a risk–benefit assessment is made, the size of benefit can be very difficult to estimate based on a surrogate endpoint. Biomarkers cannot serve as final proof of clinical efficacy or long-term benefit. If they are intended to be the basis for regulatory review and approval then, unless they are properly validated, there should be a predetermined plan to supplement such studies with further evidence to support clinical benefit, safety and risk/benefit assessment.

5. Choice of Control Groups

Ideally, we wish to obtain an unbiased estimate of the effect of the treatment being investigated compared to placebo or to another active compound and, for this reason, every effort should be made

to randomise patients from the beginning of the therapeutic testing phase. The goal of obtaining an unbiased estimate of the size of effect is true in studies in small populations as well as large trials for common diseases. Thus, in developing any treatment, a comparative trial will usually be *preferable* but may not always be *possible*. In serious and life-threatening diseases where no alternative treatments exist, there can be a tendency to grasp at any treatment seemingly offering some hope to patients. Anecdotal reports of patients responding then make it practically very difficult to persuade patients to take part in subsequent controlled trials.

In general, there are two approaches to selecting control patients: internal controls or external controls, who may be historical or concurrent. The ideal is a comparative trial using an internal control group, as there are several well-known problems inherent with historical (or other external) controls.

If there are any strong prognostic factors for the outcome, then a stratified randomisation procedure, combined with a suitably stratified/modelled analysis can greatly increase the efficiency of the trial. Similarly, such stratification – across as many factors as possible – will usually increase credibility of the results by ensuring balance on these factors across the treatment groups. Stratification for many factors in small studies becomes almost impossible unless dynamic/covariate-adaptive randomisation schemes are used.

Although internal controls are the preferred option for comparative trials, under exceptional circumstances external controls may be acceptable. Historical controls (using patients treated with ‘current’ therapies, or not treated at all) might, under exceptional circumstances, be acceptable to demonstrate efficacy, safety, ease of administration and so on, of a new treatment. In general, the absence of any control data is only likely to be acceptable if the natural course of the disease is very well known.

Patient registers may supply important information on the natural course of disease and may help in the assessment of effectiveness and safety. Further, such registers might be used as a source for historical controls.

If only active controlled studies are possible, then showing equivalence or non-inferiority may be difficult because assay sensitivity of the study cannot be assured and so obtaining a licence in these circumstances becomes extremely difficult. Arguments concerning the natural course of a disease may help to support assay sensitivity of studies.

6. Methodological and statistical considerations

Even though there are no statistical methods particularly intended for small samples there do exist methodological applications that might be helpful in this context. The following text discusses a range of approaches, which may be helpful in particular situations. As already mentioned at the beginning of the document, any given list of possible approaches cannot be exhaustive and this list is not intended to be so.

Each approach – not only those listed here – has to be weighed according to its merits and drawbacks on a case-by-case basis. Sponsors are encouraged to seek scientific advice or protocol assistance when considering use of alternative designs and methodologies.

6.1 Design stage

In conventional phase III trials sponsors are expected to enrol several hundred or even thousands of participants. The design and conduct of any trial should be such that a minimum of bio-noise is produced. Bio-noise is the sum of avoidable and unavoidable non-systematic errors in the design and conduct of a trial. It usually (although not always) leads to a bias towards failing to show a difference between treatments. As an example, a typical error, which has avoidable and unavoidable elements, is loss-to-follow up. Researchers cannot force patients to stay in a study but there is empirical evidence that some measures may help to reduce the loss-to-follow-up rate in longitudinal studies. Examples include ensuring visits are scheduled at reasonable intervals and at times convenient to patients, providing transport for patients where necessary, etc.

While in a large trial the impact of this noise-to-effect ratio can usually be reduced simply by increasing the sample size, it can become a severe problem in small studies. Therefore it is of utmost

importance that sponsors pay great attention to the minimisation of avoidable errors. Important considerations are given in the ICH guideline on the statistical principles for clinical trials (E9) as well as to other standard texts on the design and analysis of clinical trials. Many of the following methods are likely to reduce the amount of bio-noise and thus increase the efficiency of a study – but in nearly all cases at a cost of increased complexity and possibly also bias.

Continuous variables usually allow for higher precision/smaller sample size than those that have been categorised or even split into ‘responder’ vs. ‘non-responder’. This is particularly true if the baseline value is accounted for in an appropriately pre-specified analysis of covariance (ANCOVA) model. Even when baseline measurements of the eventual outcome variable may not be available, other important prognostic variables are likely to increase the efficiency of an ANOVA or ANCOVA analysis and will almost never decrease its efficiency. Unreliability of one particular outcome can also be avoided by choosing another outcome (as long as this outcome is clinically meaningful and prior to the study starting), training of outcome assessors, and using multiple ratings. All of these aspects should be considered before the study starts. Careful choice of the most *efficient* endpoint (even if not the most clinically relevant) will help to demonstrate the effectiveness of treatments. However, the size of the clinical effect is always necessary when considering the balance of risks and benefits.

Randomisation procedures

Matching or stratification (including so-called ‘minimisation’ procedures) also improve power, particularly if matching or stratification is based on important prognostic variables. Note, in addition, that in real, practical cases, these approaches rarely *reduce* study power by any important amount (even if the stratification factors turn out to be unimportant). Such procedures, accompanied by pre-specified stratified analyses and sensitivity analyses may, therefore, be useful.

Covariate-adaptive methods

These methods are sometimes used instead of stratification. Generally the principle of such covariate-adaptive, or dynamic, methods (including, but not limited to ‘minimisation’) is that each new allocation may lead to imbalances between the groups with respect to measured covariates. These methods then aim to correct that imbalance, as far as is possible, by allocating the next patient to one or other of the treatment groups. This is achieved by changing the probability of being allocated to one particular group, based on the characteristics of patients already assigned, and on the characteristics of the next patient to be assigned.

The problems with such an approach are that they are not strictly ‘random’ and that conventional statistical methods cannot be used for data-analysis.

Covariate-adaptive methods are likely to be suitable when randomisation should be stratified but there are too many factors to make stratification feasible. This is particularly the case when trials are small relative to the number of stratification factors and levels. It is, however, unclear whether unknown confounders are necessarily balanced by this procedure. Further, if centre is used as a stratum and there are many centres but few patients per stratum this may lead to simple alternate allocation, thus jeopardising allocation concealment.

There is good evidence that such methods are effective at producing well-balanced groups, even in small trials and even when a trial might have to be terminated early. The stratification factors must, however, be properly accounted for in the analysis and the impact on variables not used for stratification (both those measured and unmeasured variables) should be discussed.

Response-adaptive methods

Instead of changing the allocation of patients to treatments to achieve balance of baseline covariates, response-adaptive designs change the allocation ratio based on which treatment appears to be ‘best’. As patients complete a trial, if one treatment is beginning to emerge as better, then new patients entering the study are more likely to be allocated to that treatment. These designs are sometimes called ‘play-the-winner’ designs. The allocation probabilities can be continuously changing and do not rely on ‘good evidence’ of one treatment being superior (when, possibly, a study might be terminated anyway). As soon as one treatment appears better, the allocation of new patients is biased in favour of that treatment. As the study continues, the apparently ‘best’ treatment may change and allocation bias

can change with it. Such methods rely on outcome data being available quickly (relative to patient recruitment) and also rely on continuously unblinding individual patients as they complete the study. The analysis may be very complex as it is not based on standard assumptions of equal and constant probability of being assigned to either treatment.

A variation of response-adaptive designs is those used for dose finding – they are typically referred to as ‘continual re-assessment’ methods. They are sometimes, but rarely, used. The properties of such methods far outstrip those of conventional ‘up and down’ dose finding designs. They tend to find the optimum (however defined) dose quicker, they treat more patients at the optimum dose, and they estimate the optimum dose more accurately. Such methods are encouraged during all phases of development.

Sequential designs

Sequential designs – with a goal to demonstrate ‘statistical significance’ if a treatment is genuinely superior to control - generally reduce the required sample size. There can be several different types of sequential design – all providing valid statistical conclusions but each tailored to specific balances of expected outcomes and patient availability. Some designs are ‘open-ended’ and (in theory) continue to recruit patients until a reliable positive or negative conclusion about the treatments can be made. Other designs are ‘closed’ and so have a fixed upper limit to recruitment (but may stop before this). Stopping boundaries for benefit and harm need not be symmetrical; boundaries for showing benefit of treatment relative to an active control also need not be symmetrical. Sequential designs, as with response-adaptive designs, require treatment outcomes to be available quickly (relative to the patient recruitment rate). This will almost never be the case if we are looking for long term survival data, for example, but may be the case if we are looking at shorter term clinical or surrogate/bio-markers. A common problem with trials in rare diseases is that recruitment *is* slow because patients are so rare; hence such methods may have more of a place in these situations than in more common diseases. Ultimately, however, the extent to which the sample size can be reduced depends on the size of the effect.

n-of-1 trials

In this design the unit of randomisation is the intervention, rather than the patient. They are rather like crossover studies, but carried out in single patients. The patient’s first treatment is determined at random and at the end of a treatment period, the patient is randomised again; a switch to the alternative instead of randomisation is also possible. Multiple switches may occur. The outcome of such a study is a conclusion about the best treatment *for this particular patient*. Series of *n-of-1* trials may begin to show trends for repeated preference of one treatment over another. The advantage of such a design is that each patient is assured of ‘eventually’ ending up on the treatment that is best for him or her. Different investigators can use the same treatment comparisons in ‘their own’ *n-of-1* trials and do not have to conform to a standard protocol, which may be too restrictive in some individual cases. So rather than a patient being excluded from a trial because they do not meet the inclusion criteria, or because they would not be able to follow all the necessary trial procedures, each trial can be tailored to each patient. *n-of-1* trials have many of the same limitations as cross-over trials. They are most useful for fast-acting symptomatic treatments and in diseases that quickly return to stable baseline values after treatment. Results of many *n-of-1* trials may then be combined in a manner similar to both a cross-over study and a meta-analysis.

6.2 Data analysis

Assumptions

Studies with few patients are often perceived as presenting a rather simple situation: there is not much information (data) and so simple (often descriptive) analyses are all that are warranted. It seems quite counterintuitive, therefore, that for ‘simple’ situations more complex approaches should be applied but this is exactly what is necessary. Crude (simple) methods may often be adequate when we have huge amounts of data – but when there are very few data, it is imperative that the most efficient and informative analytical methods should be used. Many of these methods involve ‘statistical modelling’. Such models usually make assumptions about the data or the form of the treatment effect. With few data, these assumptions may not be testable or verifiable. However, assumptions *add* to the

data so that more complex statistical models give us more information than simple descriptive statistics. Hence, sensitivity analyses consisting of various analyses/models should be presented, which may make different assumptions about the data. Then it can be seen if the conclusions are heavily reliant on the model assumptions or if, in fact, they are robust to a variety of plausible assumptions.

Non-parametric methods

Contrary to the above, non-parametric, or ‘distribution-free’ methods may often be used when we cannot determine if data are from a Normal (or other specified) distribution. There are a wide variety of methods that make ‘few’ (although not usually ‘no’) assumptions about the data or about the form of any treatment effect. Some forms of Bootstrap methods make no assumptions about data distributions and so can be considered a ‘safe’ option when there are too few data to test or verify model assumptions. The primary use of these methods is in the estimation of accuracy measures (such as bias and variance) for parameter estimators, and in construction of confidence intervals. They are applied in circumstances in which the form of the population from which the observed data have been drawn is unknown. They prove particularly useful where very limited sample data are available and traditional parametric modelling and analysis are difficult or unreliable. Bootstrap methods are closely related to other data resampling methods such as the jack-knife.

α and β errors

For the stage of data analysis, several methods should be used to provide corroborative or supporting analyses. Further, as α and β error limits may be constrained by sample size, emphasis should be placed on *estimation* (point estimates and confidence intervals) rather than *hypothesis testing*. 95% confidence intervals can be used to infer whether the significance test would yield $P < 0.05$, or not. This view of confidence intervals is not helpful and a move away from ‘standard’ 95% confidence intervals to some other coverage probability may be helpful in this regard. If a sponsor chooses to do this, a good prior justification should be given.

Prognostic variables

Adjustment for baseline variables may greatly improve the efficiency of an analysis. It is mandatory for proper statistical inference that factors used to stratify the randomisation in a study should be used to stratify the analysis. As has been noted above, including stratification variables in the analysis that, in fact, have very little prognostic value rarely has any detrimental effect on the analysis. Conversely, including prognostic variables in a model can greatly enhance the precision of a treatment effect. In large scale ‘conventional’ development plans, the phase II studies would usually identify which covariates are important. In cases where we may only get one study, it may be that careful modelling to determine *which* covariates and what *functional form* they take (e.g. linear, multiplicative, etc) is necessary.

Longitudinal data

Repeated measurements over time – or in different body locations – may also improve the efficiency of an analysis. A commonly encountered problem in the analysis of such data is the non-independence between observations. Non-independence occurs when data fall into groups or clusters, e.g. in different body locations or in longitudinal studies. Standard statistical methods, such as generalized linear models (GLM) cannot be applied when analyzing dependent data, since the assumption of independence between observations is violated. Neglecting dependencies in these situations can lead to false conclusions. In general, the precision of the results and thereby their significance is usually overestimated. There are different methods available to analyze clustered dependent data e.g. the **Generalized Estimated Equations** (GEE) method, Hierarchical Linear Models or Mixed-effects models. These modern statistical approaches take the correlation within subjects into account and can also allow an unequal number of observations per subject, (e.g. caused by missing values) so that valid inferences can be assured.

Bayesian methods

Bayesian methods are a further source of ‘adding assumptions’ to data. They are a way to formerly combine knowledge from previous data or prior ‘beliefs’ with data from a study. Introducing prior

beliefs is often a concern in drug regulation. However, being able to use knowledge of likely effects of drugs due to their chemical form, likeness to other existing compounds, mechanism of action, and so on, is a very valuable addition to sparse data. As with sensitivity analyses mentioned above, a variety of reasonable prior distributions should be used to combine with data from small studies to ensure that conclusions are, at least, reasonably data-dependent and not almost entirely belief-dependent.

6.3 Interpretation of the evidence

Regulators and sponsors need to be flexible in how they interpret results from small studies that may not be as well controlled as would be expected in other areas. In 1965, Bradford-Hill described criteria for determining causality in observational studies including:

- Consistency of association / coherence with existing knowledge: Is there other evidence in favour of such an effect? How good is the available evidence?
- Biological gradient: Is there a dose response effect?
- Specificity of association: Does the product lead to this particular outcome?
- Biological plausibility: Why does the observed association make sense?
- Strength of association: A large effect usually allows bias and confounding to be easily detected.

In the absence of controlled clinical trials, these criteria may be helpful. Even in the case of randomised trials (but very small ones), they may still add help.

7. Summary and Conclusions

- There are no special methods for designing, carrying out or analysing clinical trials in small populations. There are, however approaches to increase the efficiency of clinical trials. Further, some methodological approaches, not acceptable in large trials, may be considered acceptable for trials in small and very small populations. The need for statistical efficiency should be weighed against the need for clinically interpretable results.
- Guidelines (ICH, CHMP and others) relating to common diseases are also applicable to rare diseases.
- In situations where obtaining controlled evidence on the efficacy and safety of a new treatment is not possible, the regulatory assessment may accept different approaches if they ensure that the patients' interests are protected.
- Detailed knowledge of the pharmacology of a compound may help when designing studies. Pharmacology studies may help identify sources of heterogeneity in patients. Non-clinical pharmacology (which may not be constrained by patient numbers) may be particularly helpful in conditions with very few available patients.
- Surrogate endpoints may be acceptable but need to be justified as fully as possible. Their relation to clinical efficacy must be clear so that the balance of risks and benefits can be evaluated.
- Controls and comparator groups are very important. Not all studies of rare conditions can include concurrent controls. Their absence compromises the reliability of studies.
- Patient registers may supply important information on the natural course of disease and may help in the assessment of effectiveness and safety. Further, such registers can be used as a source for historical controls. Registers used in this way should contain high quality data; GCP inspection might be anticipated.
- When planned statistical (analysis) methods fail to show treatment effects, alternative approaches should be sought out (and preferably anticipated in the study protocol). Ideally several methods should be applied and interpretation is helped if the results of different methodological/statistical approaches are in agreement.

- When using strategies and methods not frequently used in trials in large populations, scientific advice/protocol assistance is available to guide sponsors as to the acceptability for later marketing.

APPENDIX

The design stage

Surrogates

Currently it seems that the primary measure is to select a surrogate marker if recruitment of a sufficient number of patient seems jeopardized or if we have to wait too long to get an answer. These are three examples to show what level of evidence the CHMP has accepted for the authorisation of drugs to treat orphan conditions.

Example 1 is imatinib mesylate, which is authorised for the treatment of chronic myeloid leukaemia. The estimated prevalence of CML in the member states of the European Union is about 0.9 per 10.000 population. Here several trials enrolling up to 532 patients were performed (1027 patients for evaluation of efficacy in total). The primary endpoint was haematological and cytogenetic response. This surrogate endpoint was subsequently shown to be associated with survival.

Example 2 is bosentan, which is authorised for the treatment of pulmonary arterial hypertension. The estimated prevalence of pulmonary arterial hypertension in the member states of the European Union is below 2 per 10.000 population. Two randomised controlled trials were performed: a small dose finding trial (n=21) and a larger pivotal trial, enrolling 213 patients. The primary endpoint was the 6 minutes walking test; a wide range of secondary endpoints was used of which some are of clinical relevance.

Example 3 is laronidase, which is authorised for the treatment of Mucopolysaccharidosis I. Mucopolysaccharidosis I is very rare with an estimated prevalence of less than 0.03 per 10.000. One randomised placebo-controlled trial was submitted for marketing authorisation, which enrolled 45 patients. The primary endpoints were (1) forced vital capacity and (2) the 6 minutes walking test. There were several secondary endpoints of which one measured disability as a dimension relevant for quality of life.

Here we have situations where the CHMP accepted surrogate endpoints considering the limited amount of available information in relation to unmet medical needs. There is a wide range of possible situations, where available biomarkers and other surrogates will not be considered suitable.

Covariate adaptive allocation methods

Here is a descriptive example for the minimisation method [Falk 2002]: 230 patients with previously untreated, non-small cell lung cancer that is locally too advanced for resection or radical radiotherapy with curative intent, with minimal thoracic symptoms, and with no indication for immediate thoracic radiotherapy were randomly allocated by using a minimisation procedure stratified by (1) clinician (at least 24 strata – exact number is not given), (2) histology (4 strata), (3) presence of metastases (2 strata), and (4) WHO performance status (4 strata), to supportive treatment plus either immediate or delayed thoracic radiotherapy. This results in $24 \times 4 \times 2 \times 4 (=768)$ strata, which is very large, even for a large study. In this example minimisation was used to ensure that groups are very well balanced and random noise does not unduly disturb the assessment of the intervention's effect.

Responsive-adaptive methods

A group of 12 newborns with respiratory failure were to be allocated to standard treatment or extracorporeal membrane oxygenation using the 'play the winner rule'. The first patient was randomly assigned to conventional treatment (this patient died); 11 patients were then chosen for extracorporeal membrane oxygenation (all survived). After the twelfth patient the trial was terminated (Bartlett, 1985). This particular trial was heavily criticised and a subsequent randomised controlled trial (O'Rourke, 1989) was also criticised as randomisation was halted after the first 4 deaths in the control group. Modifications to the play-the-winner rule, e.g. including a random element, are possible. Such an approach is probably often not a suitable way forward but there may be indications for a similar approach.

In another example, there was also an attempt to maximise the proportion of patients who receive the apparently 'best' treatment, Giles *et al.* (2003) used a response-adaptive algorithm. There were three treatment arms: idarubicin + ara-C (IA), troxacitabine + ara-C (TA), and troxacitabine + idarubicin

(TI). Initially, randomisation was in the ratio 1:1:1 but as patients were treated and responses observed, the allocation ratio was changed to favour the most promising treatment(s). Thirty-four patients were treated. After first 5 patients had received TI, this arm was stopped (the allocation ratio went to zero). When 34th patient was recruited, the allocation ratio was 0.959:0.041 in favour of IA. The study was then stopped concluding that IA was superior to either TA or TI.

Sequential designs

Trnavský *et al.* (2004) use an adaptive group-sequential design in a study for knee osteoarthritis. The study is essentially a 'standard' randomized, parallel group, double blind design but has three scheduled interim analyses (in addition to the potential 'final' analysis). At each interim analysis, as well as the potential to stop the study for efficacy, the planned study size was also re-estimated. The first interim analysis was scheduled after 12 patients per treatment group. The study would have stopped (for positive efficacy) if the observed p-value was less than 0.0041. The study did not produce such strong evidence and so it was continued but with a revised total sample size. At the second interim analysis (after 25 patients per group), the study did reach statistical significance (at the corrected significance level) and so it was stopped and concluded a benefit of ibuprofen cream over placebo.

A fully sequential design was conducted by Sharpe *et al.* (2003) in patients receiving orthotopic liver transplantation. Patients were recruited in pairs: one randomly allocated to itraconazole and the other to placebo. After each pair of patients, a 'preference' for one or other (or neither) treatment was determined based on which patient did not (or did) suffer a subsequent infection. Of the first 71 patients, 9 suffered an infection following treatment with placebo (itraconazole 'wins') and only one suffered infection following itraconazole (placebo 'wins'). The study was this able to stop at this point having shown a risk ratio of 0.24/0.04=6, P=0.04 (Fisher's exact test).

A further example of a fully sequential design is given by Cheng, Chang and Yuen (2004). In this case, the investigators used a cross-over design so that each patient could show a preference for one or other treatment (unlike the previous example where patients were randomised in pairs). The treatments were oral rinses to alleviate mucosistitis in children undergoing chemotherapy. Forty patients were recruited of whom six did not complete the study. However, following 34 patients completing the cross-over regimen, the study was stopped, having reached statistical significance ($P<0.05$) in favour of chlorhexidine.

n-of-1 trials

In a trial 51 patients with osteoarthritis who were uncertain whether non-steroidal anti-inflammatory drugs (NSAID) were "helpful" for them were randomised either to a conventional treatment group (n=25) or to an n-of-1 group (n=24) (Pope *et al.* 2004). The n-of-1 group received in a random, double blind manner NSAID or placebo for 2 weeks over a total study period of 3 months. All patients in the control group received 'conventional' treatment, i.e. NSAID. NSAID appeared to be effective in 81% of the n-of-1 group and in 79% in the "conventional" group; none in the placebo group preferred placebo. Here the n-of-1 trial was used to compare the effect observed in a placebo-controlled trial with conventional NSAID treatment. However, theoretically only 24 patients would have been necessary to determine effectiveness as compared to placebo.

Another example, also in osteoarthritis was published by Wegman *et al.* (2005). Thirteen patients were selected and randomised to five sequences of two weeks' of NSAID and two weeks' of paracetamol. Only 5 patients completed the study and little difference was seen between the two treatment regimens. Whether this is due to a lack of differential effect, or due to low power remains unclear.

The Analysis stage

Assumption free methods

A cost-effectiveness analysis is used to highlight this issue (Korthals-de Bos IBC, 2003). Costs are often severely non-normal in distribution because there are always a few patients who use a lot of resources. Korthals-de Bos *et al.* used bootstrapped estimates of costs and effectiveness to construct a convincing graph: compared with physiotherapy, manual therapy is most likely more effective and cheaper. A scenario where manual therapy is less effective while being more expensive is unlikely.

Non-parametric methods

Example: There are 100 controls (which are usually not concurrent). You can only look at 5 cases within a given period but plan to look at 5 such clusters. You want to show that a predefined number of patients have a value above the median of the controls and you define a confidence interval (e.g. 95%). Based on this you can either (a) calculate the power, or (b) analyse the results. Ad (a), Power: e.g. you want to know the power of such a study if you assume that 3 out of 5 have a value above the median of the controls; the lower confidence limit may be defined as 15 (i.e. 3 out of 5 have a value below the 15th value of the controls) and 85 (i.e. 3 out of 5 have a value above the 85th value of the controls) which is almost a 95% confidence interval. If indeed this effect were present the study then would have 80 power to detect such an effect. Ad (b): Now we have 10 controls and only two sequential clusters of 5 cases. The limits are 1 and 9 (order statistics for controls). Using these prediction limits indicates that the probability that 3 of 5 samples in each of 2 replicates are above the 9th control sample by chance is 0.018; the probability that 3 of 5 samples in each of 2 replicates are below the 1st control sample by chance is 0.003.

Relaxing the type I and II error margins

Relaxing the type I error boundary increases the risk of false positive trial results. There are however, situations where such an approach is acceptable. One example is a trial in patients with granulomatous disease, a very rare disorder with an annual death rate of 2 to 5% (Gallin, 2003). Patients suffer from severe bacterial and fungal infections. Antibiotic prophylaxis reduced the incidence of serious infections greatly. The question was whether prophylaxis with itraconazole reduces the incidence of fungal infection. The problem was that the incidence of fungal infection is about 0.1 per patient year, which makes almost impossible to recruit a sufficiently large sample. 39 patients with chronic granulomatous disease were enrolled; each year patients were newly randomised using a biased coin to assure that the number of exposures was equal in both groups. The two-sided p-value to reject the null-hypothesis was 0.102. After 12 years of follow-up 1 in 61 patient years under itraconazole occurred as compared to 7 in 63 patient years under placebo ($p=0.10$). The investigators concluded that itraconazole was effective in this indication.

Bayesian methods

Tan *et al.* (2002, 2003) show how Bayesian methods may be used in the analysis of trial data from rare diseases. Existing information is weighted according to validity and how relevant these trials are to the question under study. These data then can either be used to estimate the necessary sample size or to inform the data gained from a small trial. This approach is related with the empirical Bayes approach, where data are informed exclusively by (unweighted) study data. Further and they describe how to use predefined scenarios within the frame of a sensitivity analysis: they suggest a sceptical prior distribution (it is assumed that the new therapy is even worse than standard treatment; a neutral prior distribution (the new therapy has no effect at all); an enthusiastic prior distribution (the new treatment has a predefined realistic effect). In this example the scenarios are used to inform a data-monitoring committee whether a proposed trial could provide useful results. This approach clearly can be used at later stages of a trial as well.

Bibliography

Bartlett *et al.* Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Paediatrics*. 1985;76:479–487.

Cheng KKF, Chang AM, Yuen MP. Prevention of oral mucositis in paediatric patients treated with chemotherapy: a randomised crossover trial comparing two protocols of oral care. *Eu J Cancer* 2004;40:1208–1216.

Gallin JI *et al.* Itraconazole to Prevent Fungal Infections in Chronic Granulomatous Disease. *N Engl J Med* 2003; 348:2416–2422.

Giles FJ, Kantarjin HM, Cortes JE, Garcia-Manero G, Verstovsek S, Fadel S, Thomas DA, Ferrajoli A, O'Brien S, Wathen JK, Xiao LC, Berry DA, Estey EH. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *J Clin Oncology* 2003;21:1722–1727.

Korthals-de Bos IBC, *et al.* Cost effectiveness of physiotherapy, manual therapy, and general practitioner care for neck pain: economic evaluation alongside a randomised controlled trial. *BMJ* 2003;326:911.

O'Rourke PP, Crone RK, Vacanti JP, Ware JH, Lillehei CW, Parad RB, Epstein MF. Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomized study. *Pediatrics*. 1989;84(6):957–963.

Pope JE, Prashker M, Anderson J. The efficacy and cost-effectiveness of n of 1 studies with diclofenac compared to standard treatment with nonsteroidal antiinflammatory drugs in osteoarthritis. *J Rheumatol* 2004;31:140–149.

Sharpe MD, Ghent C, Grant D, Horbay GLA, McDougal J, Colby WD. Efficacy and safety of itraconazole prophylaxis for fungal infections after orthotopic liver transplantation: a prospective, randomized, double-blind study. *Transplantation* 2003; 76:977–983.

Tan SB, Machin D, Tai BC, Foo KF, Tan EH. A Bayesian re-assessment of two phase II trials in gemcitabine in metastatic nasopharyngeal cancer. *Brit J Cancer* 2002;86:843–850.

Tan SB, Dear KBG, Bruzzi P, Machin D. Strategy for randomised clinical trials in rare cancers. *BMJ* 2003;327:47–49.

Trnavský K, Fischer M, Vögtle-Junkert U, Schreyger F. Efficacy and safety of 5% ibuprofen cream treatment in knee osteoarthritis. Results of randomized, double-blind, placebo-controlled study. *J Rheumatol* 2004;31:565–572.

Wegman ACM, van der Windt DAWM, de Haan M, Devillé WLJM, Fo CTCA, de Vries Th PGM. Switching from NSAIDs to paracetamol: a series of n of 1 trials for individual patients with osteoarthritis. *Ann Rheum Dis* 2003;62:1156–1161.