Clinical trials of antibacterial agents: a practical guide to design and analysis†

Caroline Smitha, Carol Burleyb,*, Mick Iresona, Tony Johnsond, Debbie Jordanf, Sidonie Knightf, Trevor Masona, Dan Masseyh, Julie Mossf and Keith Williamsa

a Zeneca Pharmaceuticals, Alderley Park, Macclesfield, Cheshire SK10 4TG; b Hoechst Marion Roussel, Broadwater Park, Denham, Uxbridge UB9 5HP; c SmithKline Beecham, New Frontiers Science Park South, Harlow CM19 5AW; d MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR; e ClinTrials Research Ltd, Kings Chase, 107/123 King Street, Maidenhead SL1 1DP; f Glaxo Wellcome, Greenford Road, Greenford UB6 0HE; g Pfizer Central Research, Sandwich, Kent CT13 9NJ; h BIOS (Consultancy and Contract Research) Ltd, Pinewood, College Ride, Bagshot, Surrey GU19 5ER, UK

Guidelines on the conduct of clinical trials of antibacterial agents produced by the US Food and Drug Administration, the British Society for Antimicrobial Chemotherapy, the Infectious Diseases Society of America and a European Working Party have been reviewed. Although very informative, these guidelines provide limited practical guidance on the design and statistical aspects of phase III studies of antimicrobial agents. This paper describes the differences between antibacterial trials and clinical studies in other therapeutic areas with regard to subjective endpoints, dual clinical and bacteriological endpoints, frequent protocol violations and difficulty of using placebo controls. The importance of a detailed protocol and planned analysis strategy is emphasized. The choice of comparator agents, practical issues with the blinding of trial materials and the documentation of patients excluded from study entry are discussed. The use of different patient groups and different endpoints in analyses are described. The principles of equivalence and their application to trials of antibacterial agents are discussed, together with an approach to calculating sample size. A variety of statistical analyses of results are compared for different situations indicating some of the problems that can arise. Different methods of presentation of study data are included with emphasis on regulatory submissions rather than scientific publications. Some graphical presentations are recommended and issues regarding data across different studies are discussed.

Introduction

The first set of guidelines for the conduct of clinical trials with antibacterial agents was published by the Food and Drug Administration (FDA) in 1977. However, these were only outline principles and the increased complexity of clinical trials necessitated the updating of these guidelines. The British Society for Antimicrobial Chemotherapy (BSAC) published guidelines in 1989. Both the Infectious Diseases Society of America (IDSA), under a contract with the FDA, and a European Working Party produced general and disease-specific guidelines in 1992 and 1993, respectively. The IDSA guidelines were later supplemented by two additional publications.

The guidelines published to date incorporate advice on general aspects of ideal trial design and analysis, but specific recommendations regarding the practical conduct and interpretation of clinical trials of antibacterial agents, especially with respect to statistical issues, are lacking. In particular, it is uncommon that all patients treated follow the protocol, leading to the need for interpretation of entry criteria and outcomes. Against this background, PSI

†Report of a PSI (Statisticians in the Pharmaceutical Industry) Working Party

*Corresponding author.

© 1998 The British Society for Antimicrobial Chemotherapy
and advice on data presentation. A glossary of terms is considered to address where they are difficult to implement as well as issues which they do not address.

This paper, which represents the conclusions of the Working Party, focuses on comparative clinical trials of antibacterial agents in the treatment of acute infections. Many comments also relate to trials of antifungal and antiviral agents and to trials of prophylaxis. Specific issues of antibacterial agents in trials of neutropenic patients are not addressed.

Clinical trials of antibacterial agents differ from other clinical trials in several ways, which combine to make studies in this therapeutic area particularly complex. Infection involves a three-way interaction between the antibacterial agent, the patient and the pathogen causing infection. These interactions affect the approach to the design, analysis and interpretation of results. Efficacy must be assessed in two ways, clinically and microbiologically. In addition, studies of antibacterial agents often include patients with potentially life-threatening conditions, where treatment must commence before the identity and sensitivity of the pathogen(s) is known. This leads to practical problems such as misdiagnoses, inappropriate treatment and inappropriate sampling times. The Guidelines offer limited advice on such issues and hence there can be difficulties in interpretation of data.

This paper addresses: the typical trial design of antibacterial studies; some issues relating to design; the measurement of efficacy; which patients should be included in each particular analysis; how patient numbers should be estimated; which analyses should be performed; and advice on data presentation. A glossary of terms is given in Appendix 1.

**Trial design**

The typical phase III antibacterial clinical study fulfils the following design criteria. (i) Parallel groups are used: crossover studies are inappropriate, since the underlying disease does not return to a baseline state and many patients are cured at the end of treatment. (ii) There are active controls: a placebo control is unethical when effective treatment is available. (iii) There are usually two treatment arms: the most appropriate design is to compare the new treatment against the ‘gold standard’ (a treatment regimen may be a single drug or a combination of drugs). (iv) Multiple centres are involved so that large numbers of patients can be studied within a reasonable time-frame and so that a representative sample of bacteria and susceptibilities across several locations can be obtained.

(v) The study is randomized to eliminate treatment allocation bias. (vi) A double-blind design is used to reduce assessment bias; this is practical for oral treatments, but can be impractical for studies of intravenous or intramuscular drugs. At a minimum, studies should always be assessor-blinded. (vii) Most studies are designed to show equivalence. This is discussed in further detail in the section entitled ‘Principles of equivalence’. (viii) Studies involve a single course of treatment, as this is sufficient to eliminate the majority of infections. (ix) Treatment is rarely for more than 10 days, with a follow-up of 2–4 weeks. (x) Studies are single entry, i.e. patients should only enter the study once in chronic or recurrent infections. (xi) Both clinical and microbiological assessments are performed at the end of treatment and at follow-up.

Studies may be conducted in either community or hospital in-patient settings and in a single infection, e.g. pneumonia, or a group of infections, e.g. abdominal infections, although it is preferable to subdivide these, e.g. into biliary, appendix and colorectal infections.

**Entry criteria**

The Guidelines give clear direction on entry criteria for infections at a single site, but provide no guidance for patients with infections at multiple sites, which often occur in studies of severely ill patients and in patients with septicaemia, which has often arisen from one or more local infection sites. In such cases it is essential to consider the aims of the study and to define clearly the entry criteria and how these patients will be handled in the analysis. The Working Party recommends that in a study of one particular infection, patients with additional infections at other sites may be permitted to enter the study and the information on the other sites collected and summarized, but this should not contribute to the assessment of the efficacy of the drug at the site of interest. It is recognized that there may be confounding factors and a balance is needed between the ideal and the real situation. Specific studies can be set up to investigate the effect of the drug in patients with more than one infection. Methods of analysis of these studies are outside the scope of this paper, although it is important to remember that the principle of independence of observations must be upheld, i.e. each patient should only contribute once to any analysis.

One area where the Guidelines are not considered by the Working Party to be sufficiently robust concerns the inclusion of patients who have already received antibacterial agents to treat the current episode of infection. The most rigorous approach is to disallow entry of such patients; this is practical in some studies, e.g. community-acquired pneumonia in general practice, but impractical for others, e.g. meningitis. The protocol should state the duration and timing of previous therapies which...
are permitted, based on the potency, spectrum and half-life of each therapy. The Working Party recommends that patients who have taken previous therapy may be included provided that the pathogens present are known to be resistant to that therapy or that clinical signs and symptoms remain, i.e. in clear cases of failure of the previous therapy. Decisions taken on the handling of patients who have taken previous therapy in the analysis must be clearly documented and justified.

Exclusion criteria should be kept to a minimum and identify patients who are at greatest risk of suffering adverse effects from the trial or control drugs. Patients should be excluded if any pathogens isolated are known to be resistant to either study treatment. Often the identity and antibiotic susceptibility of pathogens are not known at the start of the trial and it is essential to allow patients to be withdrawn from the trial if pathogens are subsequently found to be resistant and the patient is not responding.

**Patient log**

Some guidelines recommend that a log is kept of all patients considered for inclusion in a trial. This can provide some evidence that entry of patients has not been biased by knowledge of the treatments, particularly with regard to patients excluded because of unfavourable expectation of one treatment arm. The log can also quantify exclusions due to organisms resistant to only one of the treatments. It is recommended that all patients with the infection of interest are included in the log, which is kept simple, e.g. records of age, sex, nature of infection, reason for exclusion from study. There can be practical difficulties in ensuring that such a log is complete, but the log is an important part of the trial documentation and the data should be summarized in the final trial report. As these patients have not consented to the study (indeed, non-consent may be the reason for exclusion), the data on the patient log cannot be verified against patients’ notes like study data. Although this weakens the usefulness of the patient log, it is still considered valuable.

**Choice of control drug**

The control drug regimen should be widely accepted for the particular infection and may differ between indications so that the full spectrum of activity of the new drug can be explored. Accepted treatments, doses and formulations may differ between countries, because of different licensing or prescribing policies, determined by local resistance patterns and cost considerations. The Working Party recommends that only one comparator treatment regimen is used in any one trial, even if this limits the number of participating countries. Justification of the choice of control drug and dose should be given in the protocol.

**Efficacy**

The efficacy of an antibacterial agent should be assessed clinically and also microbiologically, as pathogens can be eradicated without clinical cure and vice versa. All patients can be assessed clinically, but it may not be possible to assess all microbiologically. Assessments are usually made both at the end of treatment and also after a suitable follow-up period. The follow-up assessment is used to detect relapses, which may indicate that the infection was only suppressed, rather than eliminated, at the end of treatment. Responses are categorical, and clinical and microbiological responses must fit only one category.

The Guidelines recommend an overall outcome assessment which combines clinical and microbiological assessments. The Working Party believes that this is unnecessary, as in practice either the clinical or the microbiological response is designated as the primary endpoint.

**Clinical response**

The clinical response is an assessment of the change in the patient's original infection. It is based on objective measures such as signs and symptoms of disease evaluated before, during and after treatment. These are well discussed in the Guidelines. The Guidelines recommend use of the terms 'cure' and 'failure' for assessing response, but suggest that others, such as 'improved', can be used if defined in the protocol. At the analysis stage, categories are usually combined to provide a binary response of success or failure and thus any other category is rarely useful as it adds little to the overall assessment of treatment. The Working Party recommends that response is always recorded as either 'success' or 'failure', where the definitions of success and failure may vary between infections (Table I). However, it should be made clear how a patient who has 'improved' should be classified, e.g. all symptoms responding such that the patient can be discharged. A response of 'indeterminate' should only be used when a patient cannot be assessed, e.g. lost to follow-up.

A relapse is defined as a patient with a successful outcome at the end of treatment, but who subsequently fails before or at follow-up, i.e. has a return of signs and/or symptoms. A separate response of relapse need not be recorded in the Case Report Form as it may be derived from the assessments at the end of treatment and follow-up. Microbiology will indicate whether the relapse is due to re-infection with the same pathogen or super-infection.

If the patient develops an infection at a different site, this should not contribute to the clinical response. This information must be recorded separately; one approach is to record it as an adverse event.
Microbiological response

A bacterial cause of infection, i.e. a pathogen, should be identified from an appropriate sample at the start of the study whenever possible. The responsibility of recording organisms as pathogens should lie with the investigator or microbiologist. Other data which should be recorded are: date of collection, specimen type, quantitative evaluation of pathogens for certain infections (e.g. urinary tract infections) and susceptibility to trial treatments (e.g. zone sizes or MICs).

Accurate and consistent susceptibility testing of the pathogens is essential. This may be achieved by use of central laboratories, which have the advantages of standardized techniques and equipment. However, there are also disadvantages, e.g. loss of ‘fragile’ organisms in transit and possible duplication of testing if also carried out at a local laboratory (for immediate patient care), leading to discrepancies in identification or sensitivity.

For each pathogen isolated before treatment, a response should be given at the end of treatment and at follow-up. This responsibility should lie with the sponsor, using clearly defined rules documented before trial initiation. These responses are detailed in the Guidelines and the main ones are summarized in Table II.

New organisms may be isolated during treatment or within a specified time window after treatment has stopped. If the organism is considered a pathogen requiring treatment, then the response is superinfection; if the organism does not require treatment then the response is colonization.

Assigning microbiological responses is not always straightforward. Two examples of patient outcome are given in Table III, with the appropriate responses at the end of treatment. Table III emphasizes that Enterococcus faecalis and Klebsiella pneumoniae were eradicated rather than showing that the patient failed in some unspecified way.

‘By patient’ microbiological response

One assessment not mentioned in the Guidelines is an overall patient response based on microbiological responses for each pathogen. This response summarizes the microbiological activity of the drug across pathogens and is most useful when patients have more than one pathogen, as in intra-abdominal and gynaecological infections.

The ‘by patient’ microbiological response can be easily derived from individual pathogen responses (Table IV). If all pathogens have the same response then the microbiological response for the patient is straightforward to define. If there is a mixture of pathogen responses, then the worst case is taken, based on the rule that if any pathogen ‘persisted’ then the patient response is failure. This patient response can be statistically analysed to assess the microbiological effect of the treatments across patients. In such an analysis, the responses are divided into satisfactory and unsatisfactory categories (Table IV). The Working Party strongly recommends the use of this ‘by patient’ response.

Table I. Example of definitions of clinical responses

<table>
<thead>
<tr>
<th>Response</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>resolution of signs and symptoms of original infection or, especially in chronic cases (e.g. cystic fibrosis), a defined improvement in signs and symptoms</td>
</tr>
<tr>
<td>Failure</td>
<td>insufficient improvement or a deterioration in signs and symptoms of infection; death due to infection; further antibacterial agents required because original infection is not responding to treatment</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>patient data unavailable, e.g. patient lost to follow-up</td>
</tr>
</tbody>
</table>

Table II. Definitions of pathogen responses

<table>
<thead>
<tr>
<th>Response</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eradication</td>
<td>pathogen absent post-treatment</td>
</tr>
<tr>
<td>Presumed eradication</td>
<td>specimen unobtainable for clinical reasons and clinical response of success e.g. lack of sputum production due to success</td>
</tr>
<tr>
<td>Persisted</td>
<td>pathogen still present</td>
</tr>
<tr>
<td>Presumed persisted</td>
<td>specimen unobtainable for clinical reasons and clinical response of failure</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>specimen unavailable for other reasons e.g. patient lost to follow-up</td>
</tr>
</tbody>
</table>
Clinical, microbiological and ‘by patient’ microbiological responses are all important. The clinical response is a subjective measure indicating whether signs and symptoms are improving. The pathogen microbiological response is an important measure of the efficacy of the drug against specific pathogens. The ‘by patient’ microbiological response assesses whether the patient was pathogen-free at the end of therapy: this applies to both original and any superinfecting pathogens and therefore gives a better correlation with the clinical outcome.

Further issues relating to efficacy
It is recommended that, for all endpoints, clear-cut responses are collected whenever possible and responses classed as ‘indeterminate’ only in exceptional cases, independently of whether the patient should be included in an analysis. This makes the analysis more robust, and avoids the difficulties of investigators deciding on evaluable patients; this should be the responsibility of the sponsor, in order to ensure consistency.

The Guidelines recommend that a response is given separately at the end of treatment and follow-up for each patient and pathogen, and the Working Party endorses this approach as it avoids unnecessary assumptions. The end-of-treatment response assesses the effectiveness of the treatment to eliminate the infection, while the follow-up response assesses the potential for relapses and hence has a different function. A ‘last value carried forward’ approach can be very misleading, especially if a response of success at the end of treatment is carried forward to follow-up; this assumes that the patient has not relapsed and this may not be true.

There is often confusion in assigning a response to patients who have taken additional antimicrobial agents during a trial. If the antimicrobial agents were given to treat the original infection, which was failing to respond, then the clinical response should be failure and the microbiological response should be determined on the basis of the culture obtained just before administration of the additional therapy. In the absence of this, the microbiological response should be classified as ‘presumed persisted’. If the antimicrobial agents were given to treat a different infection and are not expected to act on the original infection, e.g. an antifungal agent was given, then the patient can be assessed for the original infection. If the antimicrobial agent was given for another infection, but it is considered that it could affect the efficacy of the trial treatment, then the clinical response should be classed as ‘indeterminate’, unless there is microbiological evidence to suggest otherwise.

If the patient dies, whether from an infection or not, then the last assessment before death should be used; this will often be failure unless the patient died suddenly with the infection clearly resolved. If a patient is withdrawn early from the study, then an assessment should be made at the time of withdrawal.

Patient groups for analysis
All patients exposed to either trial drug should be included in the assessment of safety. In the assessment of efficacy, the aim should be to analyse the comparative efficacy of the treatments whilst minimizing the potential for bias. The conventional method is to consider intention-to-treat

| Table III. Examples of microbiological response at end of treatment for a UTI study |
|---------------------------------|---------------------------------|---------------------------------|
| **Patient no.** | **Pathogen(s) at entry** | **Pathogen at end of treatment** | **Microbiological response at end of treatment** |
| 1 | E. faecalis | Pseudomonas aeruginosa (needing treatment) | E. faecalis eradicated, P. aeruginosa superinfection |
| 2 | E. coli and K. pneumoniae | E. coli | E. coli persisted, K. pneumoniae eradicated |

| Table IV. ‘By patient’ microbiological response |
|---------------------------------|---------------------------------|
| **Pathogen response** | **‘By patient’ response** | **Response in analysis** |
| All eradicated | success | satisfactory |
| All presumed eradicated | presumed success | satisfactory |
| One or more persisted | failure | unsatisfactory |
| One or more presumed persisted | presumed failure | unsatisfactory |
| Superinfection and all pre-treatment pathogens eradicated | superinfection | unsatisfactory |
C. Smith et al.

The patients included in the bacteriological PP group will be a subset of the clinical ITT group, which in turn will be a subset of the clinical ITT group. If the clinical PP group requires patients to have microbiologically documented evidence of an infection, then this group will be identical to the bacteriological PP group.

There are many slight variations of the basic ITT analysis as discussed by Gillings & Koch. Traditional approaches indicate that it should include all patients randomized, in the treatment group to which they were randomized. Variations include analysing only patients who received some treatment and analysing patients as treated, not as randomized. The merit of these different approaches is not a subject for this paper. However, in addition to patients misrandomized, there are two issues which arise more often in studies of antibacterial agents than in other areas of medicine, namely patients randomized but not treated and patients misdiagnosed. How these are handled in the ITT analysis should be defined in the protocol and the principle is that they are unlikely to affect the outcome or interpretation of an analysis unless there is a large proportion of them or they are unbalanced between the treatment groups.

The PP group of patients must be defined before the blinding of the study is broken. The purest view is to exclude all patients who do not meet the entry criteria of the trial, i.e. protocol violations, but a more pragmatic approach is to exclude only those patients who fail on criteria thought to affect treatment outcome. For patients who deviate from the protocol during the trial, rules should be agreed on what constitutes a deviation serious enough to warrant exclusion.

In studies designed to show superiority of one drug over another, the ITT analysis is usually more conservative than the PP one, as the non-compliers included in an ITT analysis tend to dilute the overall treatment effect. In an equivalence trial, the ITT analysis is no longer conservative and the PP analysis should be considered the primary analysis. However, it is clearly preferable to show equivalence in both the ITT and PP analyses and if the analyses give differing conclusions, then the data need to be carefully examined and reasons for differences proposed. One reason why the analyses may differ is because of resistance; the ITT analysis will include patients with resistant pathogens and the PP analysis excludes them.

At a minimum, the analyses should be performed at the end of treatment. At a follow-up, there are often a large number of patients without a response as they are lost to follow-up and in such situations a statistical analysis will not be informative.

**Principles of equivalence**

It would not be ethical to conduct a trial of antibacterial agents if the expectation of success was less for one treatment than for the other. Patients thought to be at different risks of failure with the treatments cannot be included in these randomized trials, hence patients with resistant pathogens are excluded and the study is ultimately performed on patients with susceptible pathogens. Such studies are often referred to as ‘equivalence trials’, although the aim is to demonstrate that the new treatment is as good as or better than the control treatment.

The principles of equivalence were addressed recently by Jones et al. The familiar statistical test and P value approach for demonstrating a difference between treatments is not appropriate for demonstrating equivalence. This is because failure to reject a statistical hypothesis does not allow the conclusion that the hypothesis is true. Therefore, testing a hypothesis that the success rates for the treatments are equal is not appropriate for demonstrating that they are equal. In any clinical trial, we cannot expect to observe or demonstrate exact equality of treatments due to random variability.

The recommended approach to both sample size estimation and analysis of the trial results is to use the

---

**Table V. Efficacy analysis groups**

<table>
<thead>
<tr>
<th>Analysis group</th>
<th>Response</th>
<th>Patient group under study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical ITT</td>
<td>clinical</td>
<td>all who meet the trial criteria. This may or may not require patients to have documented evidence of a bacterial infection, but this should be specified</td>
</tr>
<tr>
<td>Clinical PP</td>
<td>clinical</td>
<td>all who meet the trial criteria. This may or may not require patients to have documented evidence of a bacterial infection</td>
</tr>
<tr>
<td>Modified ITT (or bacteriological ITT)</td>
<td>‘by patient’ bacteriological</td>
<td>all who have documented evidence of a bacterial infection</td>
</tr>
<tr>
<td>Bacteriological PP</td>
<td>‘by patient’ bacteriological</td>
<td>all who meet the trial criteria and have documented evidence of a bacterial infection</td>
</tr>
</tbody>
</table>
observed success rates to obtain a range of values (confidence interval) within which we expect the real difference in success rates to lie. To accept equivalence, all the values in the confidence interval must be close enough to zero to be clinically acceptable. The test treatment is accepted as being as good as or better than the control provided that the lower limit of the confidence interval (test minus control) does not fall below a certain level, typically -10%, i.e. there is statistical evidence that the test treatment has a success rate at worst 10% lower than the control.

Sample size

Sufficient patients must be included to give a high probability that clear conclusions will result. Choice of sample size depends on the following parameters:

(i) Expected success rate. This is the expected success rate for both test and control treatments based on an assumption of equivalence. There are likely to be data on the active control but consideration must be given to its relevance to the current trial design. Changing resistance and doses, for example, can influence success rates and recent data are likely to be the most useful. Estimates from publications can be affected by publication bias, e.g. negative trial results are often not published.\textsuperscript{11,12} With the other parameters (ii) to (v) fixed, increasingly higher success rates over 50% require smaller sample sizes.

(ii) Maximum allowable clinical difference for equivalence. This is the acceptable difference between treatments and is interpreted as a worst-case cut-off. Where expected success rates cannot be measured with high precision, regulatory guidelines\textsuperscript{13} have indicated that larger allowable differences may be used. On the other hand, some trials might demand that equivalence be demonstrated using less than the typical 10% allowable difference. With other parameters fixed, the smaller the allowable difference the larger the sample size needed.

(iii) Confidence level. The confidence level (C\textsubscript{lev}) is the probability that the confidence interval obtained in a trial will contain the true value that is being estimated. It is recommended that a 95% C\textsubscript{lev} be used because this provides an estimate at the typical C\textsubscript{lev} while assuring a low risk (<2.5%) of wrongly claiming equivalence if in reality treatments differ by more than the allowable difference (e.g. 10%). With other parameters fixed, the sample size will be reduced by decreasing the C\textsubscript{lev}.

(iv) Probability (power) of demonstrating equivalence. This is analogous to power in statistical significance testing. It is the chance (e.g. 90%) of getting the desired result, i.e. it is the probability of acceptance of equivalence if the true success rates are equal. The complementary probability (10%) represents the risk that the trial will be inconclusive even if in reality the treatments are equal.

Ethical considerations demand a high chance of running a conclusive trial but cost considerations often require compromise. Costs involved in adding patients to raise the trial’s chance of success from 80% to 90% must be balanced against the costs of failed (inconclusive) studies. It is recommended to plan the chances of trial success at no less than 90%. As power increases, the requisite sample size increases.

(v) Evaluability rate. This is an estimate of the number of patients available for analysis. As the PP analysis is considered primary, patient numbers will need to be increased based on the estimate of the proportion of patients who will contribute to the PP analysis. This will be based on previous experience in this area or on published data and usually ranges between 65% and 95%. With other parameters fixed, the sample size increases as the expected number of patients available for the PP analysis decreases.

Methodology

There are numerous methods available for calculating patient numbers, many of which are based on the normal approximation of binomial probability. It is recommended that the values provided by Blackwelder & Chang,\textsuperscript{14} Machin & Campbell\textsuperscript{15} and Makuch & Simon\textsuperscript{16} be used. Typical settings of the above parameters and associated sample sizes are shown in Table VI. Extended tables and complete details of underlying calculations can be found in Machin & Campbell.\textsuperscript{15}

The sample size calculation gives the number of patients required for analysis. These figures must be increased to reach the number required for recruitment to meet the expected number of evaluable patients, i.e. the numbers for the PP analysis.

Parameters (i), (ii) and (v) will be difficult to estimate precisely at the design stage. For this reason it is prudent to explore the sensitivity of the sample size calculations to changes in certain parameters.

<table>
<thead>
<tr>
<th>Expected success rate</th>
<th>Clinical difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>50%</td>
<td>2102 (1570)</td>
</tr>
<tr>
<td>60%</td>
<td>2018 (1507)</td>
</tr>
<tr>
<td>70%</td>
<td>1766 (1320)</td>
</tr>
<tr>
<td>80%</td>
<td>1346 (1006)</td>
</tr>
<tr>
<td>90%</td>
<td>757 (566)</td>
</tr>
</tbody>
</table>

Table V I. Sample size per treatment group with confidence level 95% and power 90% (80%)
Statistical analysis

The Guidelines offer limited advice on statistical analysis. The following section outlines the Working Party’s recommendations for methods of analysis and when they should be used.

Features of trial design and data

In multicentre studies, patient randomization is almost always stratified by centre and possibly other factors. In many trials there will be potential confounding risk factors (e.g. age, or severity of infection), specified at the design stage, which will require further investigation. In some trials analysis may be extended to explore other factors associated with treatment response. It is also necessary to identify any factors associated with differences in treatment response (treatment-covariate interactions).

Response rates of zero or 100% within centres may cause complications with some methods of analysis. These are more likely to occur in centres with small numbers of patients.

Information collected and analysis endpoints

It is recommended that formal statistical analysis of the clinical response and of the ‘by patient’ microbiological response is performed. A formal statistical analysis of the pathogen microbiological response is not recommended as there may be multiple outcomes for each patient, violating the assumption of independent observations.

Patients with an indeterminate response will be excluded from all analysis populations except ITT, where they should be classified as unsatisfactory, as this assumes a ‘worst-case’ outcome for those with unknown response. Therefore, regardless of the quantity of efficacy data collected, all analyses will be based upon a single binary outcome.

Analysis output

The results from clinical trials should concentrate upon clinically useful information in the form of an estimate of comparative efficacy, e.g. a difference in response rates or an odds ratio and an estimate of the precision with which it has been estimated, in the form of a confidence interval. In trials set up to establish the equivalence of treatments, or at least to rule out a clinically relevant difference, P values play no part. Interpretation must be based solely upon the limits of a confidence interval and subjective judgement about where these must lie in order to establish equivalence.¹⁰

Methods of analysis

There is a vast amount of literature detailing different methods for analysing binary outcomes, as found in antibacterial trials. This paper concentrates on some of the more commonly used methods. They are listed below in order of increasing complexity.

(i) Simple comparison of proportions. This is the simplest method for analysing 2 × 2 tables and may be found in many standard textbooks.¹⁷ It provides an estimate of the difference in (percentage) response rates. A confidence interval for this difference may also be calculated, based upon a normal approximation to the binomial distribution. This approximation is less robust as the response rate for either treatment approaches zero or 100%.

(ii) Stratified comparison of proportions. The simple comparison of proportions can be extended to situations where a number of 2 × 2 tables need to be combined, for example to allow for several centres.¹⁷ A weighted average of the individual tables is calculated, where each weight is the reciprocal of the squared standard error of the difference in response rates for that table. Since the same calculations as in method (i) are applied to each table, it follows that if the observed success rate for either treatment in any stratum is zero or 100% then the variation within that stratum cannot be calculated. Such strata must either be combined with others or ignored.

(iii) Mantel–Haenszel. This method is similar to method (ii) in that it is a weighted average of estimates from several 2 × 2 tables. This time, however, it is the odds ratios that are estimated and combined.¹⁷ The estimation procedure is more robust than method (ii) with departures from normality although if, for example, both treatments have 100% success rates within a particular stratum, the odds ratio cannot be estimated. There are other possible weighting procedures for combining odds ratios.¹⁸ The recommendation is to use the Mantel–Haenszel estimate but with the estimate of variance suggested by Robins et al.¹⁹ and to check this by calculating conditional maximum likelihood estimates with exact methods for a confidence interval; this requires specialized software.

(iv) Logistic regression. The linear logistic model provides an alternative framework for estimation.²⁰ The advantage of this method is that different factors and interaction effects can be tested and allowed for very simply. The comparative efficacy of the treatments is presented by an odds ratio and a corresponding confidence interval. Model assumptions break down as success rates approach zero or 100%. If the observed success rate for either treatment in any stratum is zero or 100% then the overall estimates cannot be calculated.

(v) Exact odds ratios. Estimation of odds ratios across a number of 2 × 2 tables can be based on how extreme the observed data are relative to all other possible combinations that could have occurred. When zero or 100% success rates are observed, exact methods will at least calculate the appropriate one-sided confidence interval for the odds ratio whereas Mantel–Haenszel methods will not provide any confidence interval at all.
Design and analysis of antibacterial studies

Exact methods are available through the statistical package StatXact (Cytel Software Corporation, Cambridge, MA, USA; see also Appendix 2).

Approach to analysis

The Guidelines refer to estimates of differences in proportions and define equivalence in terms of a lower limit for the confidence interval for the difference in proportions. This implies that the analysis should be of differences in proportions, but as outlined above, this may not always be the most appropriate analysis and hence this can present the statistician with a problem. The majority of studies are multicentre and hence centre effects need to be investigated and many studies also have other stratification factors or covariates of interest, which are best investigated by other statistical methods.

The Working Party recommends that when justified, results are presented using a simple comparison of proportions (method (i)), or a stratified comparison (method (ii)), if appropriate, to take account of one stratification factor (usually centre). The robustness of this approach should be justified by investigation of stratification factors, covariates and the effects of extreme response rates by judicious use of the more sophisticated methods described in the previous section.

There is a method of back-transforming odds ratios into differences in proportions for one factor, but there is no accepted method for extending this to more than one factor and hence this method has its limitations.

Example

To demonstrate these methods of analysis, an example has been constructed (Appendix 3). Results are summarized in Table VII. Generally, results should be presented at most to two decimal places; to illustrate the differences between methods three decimal places have been used below.

None of the methods allow separate inclusion of centre since there were no failures on either treatment. In the adjusted analyses, all of the associated tests for homogeneity approached significance, suggesting the possible presence of a centre by treatment interaction.

On the evidence of the results above, there is insufficient evidence to conclude equivalence. The lower limit of the confidence interval for the difference in proportions is below -10%. The confidence intervals for the odds ratio do not contain 1, indicating non-equivalence.

Presentation of data

The following section provides advice on presenting data from antibacterial trials; no guidance is given in the Guidelines. General information on presenting data from clinical trials is provided in the ICH Guideline.

The presentation of information from clinical trials may take a variety of forms, including a narrative textual description, tabulations and the use of different types of diagrams or figures such as bar charts, histograms and graphs. The main purpose of these is to provide both a summary of the characteristics of the population to which

<table>
<thead>
<tr>
<th>Table VII. Estimates of comparative efficacy (and 95% confidence level (CI) for several methods of analysis of example in Appendix 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Difference in proportions</td>
</tr>
<tr>
<td>Odds ratio</td>
</tr>
<tr>
<td>not adjusted for centre</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>adjusted for centre</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(i) centre 9 combined</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Exact</td>
</tr>
</tbody>
</table>
the results of the trial can be applied, as well as a simple overview of results in terms of comparative estimates of efficacy and other outcomes. In clinical trial reports, these will be supported by listings of individual patient data, which will be sufficiently detailed to allow verification of summary information presented elsewhere.

Summarizing efficacy data

A clear account of all patients who entered the trial should be made, indicating how many patients belong to each analysis group. This is best represented in the form of a flow diagram (Figure 1).

Summaries of efficacy data should account for all the patients who entered the trial. All tables should clearly indicate which patient group is being summarized and present the number of patients in each cell of the table along with the corresponding percentage. Denominators used to calculate percentages should also be presented. This is particularly important at follow-up and a tree diagram may also be useful to indicate the fate of all patients during the trial. Summary tables of efficacy data should be presented by and across all potentially prognostic variables and, where appropriate, for each centre and across all centres.

Presenting the results of analysis

For each patient group, the observed response rates for each treatment should be presented (usually in the form of percentages) together with the difference between the treatments and, where appropriate, an odds ratio. An indication of precision, in the form of a confidence interval, should also be displayed. For a comparison of a number of treatment differences (or odds ratios), with corresponding confidence intervals, graphs may be used (Figure 2).

Data listings

These should provide a comprehensive record of all patient information recorded during the trial. In addition, there should be a tabular listing of all patients, visits and

![Figure 1. Flow diagram of patient groups for analysis. Percentages are calculated from the total number of patients randomized to each treatment.](image-url)
Design and analysis of antibacterial studies

It is helpful to indicate on patient data listings which patients are included in each of the analyses. Consistency of approach and attention to detail will greatly enhance the appearance of listings.

Combining data across studies for regulatory submission

When a regulatory submission is made, data from different studies will need to be combined and summarized to give an objective overview of the performance of the drug in terms of efficacy and safety. The Guidelines do not provide any guidance but some information is available in the FDA guidelines for the format and content of the clinical and statistical sections of NDA and the CPMP note for guidance.

There may be studies in several different indications, doses, routes of administration, control drugs and entry criteria. All these factors need to be considered carefully before data are combined, as inappropriate summaries or statistical analyses can introduce bias or give uninformative or potentially misleading results. It is essential, therefore, to obtain both clinical and statistical input into the decisions on when to and how to combine data and how to summarize and present the results.

Many of the problems of combining data can be avoided if careful consideration is given to the design of studies contributing to a clinical programme, such as patient entry criteria, methodology and terminology. Protocols and CRFs should be standardized wherever possible. In addition, consideration should be given at the outset to what overview summaries of data will be required, so that the results in individual trial reports can be presented consistently and in a manner which makes the overviews easy to construct and understand. This has the further advantage of providing an audit of where the data came from.

The pathogen microbiological response allows an appraisal of how well a new antibacterial agent performs against a given pathogen by combining studies in the same indication versus different comparators or across indications but taking into account any differences in dose.

The overview should be structured to meet the proposed labelling and advertising claims. Results of each individual trial should be presented, ensuring that those in similar indications using similar control drugs are presented together. This can be supplemented by graphical presentations of the results.

Conclusions

The Working Party has considered fundamental practical issues not addressed in previous guidelines and has made recommendations on how these issues should be resolved. Problems arising in the design of studies have been discussed, with particular regard to entry of patients with infections at multiple sites and patients who have taken previous antimicrobial therapy. The Working Party recommends that the clinical response is collected as a binary response of success or failure and a response of indeterminate is used only in exceptional circumstances. The use of a “by patient” microbiological response is not discussed in any of the guidelines and the Working Party emphasizes the usefulness of this response. Patient groups for analysis are not clearly stated in any guidelines and the Working Party has attempted to clarify which analyses should be performed and on which endpoints and has suggested a flowchart to assist this. Although there are many methods of analysis for binary data in the literature, there is little information available on handling the specific problems that arise in analysing data from studies of antibacterial agents. The Working Party has stressed some of the problems that arise and how these can be addressed. Debate still exists on a number of statistical points; for example, there is no accepted agreement on the primary analysis in studies of antibacterial agents. The Working Party has attempted to clarify these issues. Many problems encountered in antibacterial studies can be avoided if they are considered carefully before study initiation, are detailed in the protocol and are addressed consistently across a clinical program. Due to the complexity of studies in this area, it is not possible to prescribe solutions for every possible situation. The paper aims to highlight the important issues which must be considered during design and analysis. There is clearly scope for further work in this area.

Acknowledgements

The Working Party acknowledges the help and advice provided by Jane Hannington, Richard Bax and Paul Loughlin.
References


Received 7 May 1997; accepted 21 November 1997

Appendix 1: Glossary

Active control a pharmacologically or medically active treatment used for comparison with the test treatment.

Antibacterial (drug) a substance, either naturally occurring or chemically derived, that kills or inhibits the growth of one or more species of bacteria (in contrast, an antibiotic is a substance that is produced by or derived from a microorganism and is active against other microorganisms).

Antimicrobial (drug) a substance that kills or inhibits the growth of one or more species of microorganism (bacteria, viruses, fungi or protozoa).

Eligible (of a patient) satisfying the inclusion and exclusion criteria of the study protocol.

Entry criteria criteria which define the patients to be studied in a trial; inclusion criteria define the type of patients who should be included and are usually based on those patients who are expected to benefit from the trial treatments; exclusion criteria define those who should be excluded, usually for safety reasons.

Evaluatable (of a patient) providing data that lead to an evaluation of the effect of the study drug. The term is usually defined in terms of a patient satisfying the protocol to a well defined degree.

Intention-to-treat (ITT) traditionally, this covers all patients who should be included and are usually based on those patients who are expected to benefit from the trial treatments; exclusion criteria define those who should be excluded, usually for safety reasons.

478
patients who were entered (randomized) to the study. There are however, variations to this approach whereby data from patients who were misdiagnosed or received no study medication are not used and data from misrandomized patients is moved to the treatment group of the study medication that they actually took.

**MIC** minimum inhibitory concentration, often quoted as the MIC_{50} (the minimum concentration of an antimicrobial drug that inhibits the growth of 50% of the organisms under test) or MIC_{90} (the minimum concentration that inhibits the growth of 90% of organisms under test)

**Microbiological sample** a sample of body tissue or fluid that can be used to test for the presence of one or more species of organisms to confirm infection

**Misdiagnosis** diagnosis of a patient as having one disease/condition when in fact they do not

**Misrandomization** giving one study treatment to a patient randomized to a different treatment group

**Multicentre** conducted in more than one hospital/medical centre. Each centre is administratively separate and is responsible for recruiting, treating and providing data on patients

**Organism** in this paper we use this term to mean 'microorganism'; it is often used when the microorganism has not been identified as a bacterium, virus, fungus or protozoan

**Parallel group** a study design in which patients are randomly allocated to receive one of a number of treatments

**Pathogen** a parasite, such as a bacterium, that can produce a disease

**Per protocol (PP)** the group of patients satisfying the protocol to a well-defined degree with regard to inclusion/exclusion criteria and compliance with study procedures

**Resistance** the ability of a microorganism to withstand the presence of an antimicrobial drug

**Spectrum (of activity)** (of an antimicrobial drug) the range of organisms and the degree of activity of the drug under test

**Sponsor** the organization, usually a pharmaceutical company, providing financial support and often also administrative and scientific support for a clinical trial

**Stratified** (of patients) being separated into levels for a known factor or combination of factors, e.g.: age, sex, stage of disease before randomization, to ensure a similar mix of patients in each randomized group

**Susceptibility** the level of sensitivity or resistance of a microorganism to an antimicrobial drug

**Zone size** the size (diameter) of the area of no-growth around an antimicrobial disc on an agar plate which has been seeded with a known concentration of bacteria. The zone size provides a measure of the activity of the antibacterial drug

### Appendix 2: Odds ratios

A common way of presenting estimates of comparative efficacy is through the use of odds ratios. Although less intuitively appealing than differences in response rates, they often provide a more meaningful basis for statistical analysis and certainly a better framework for statistical modelling. This is especially the case when response rates approach the extremes of zero and 100 and where calculation of confidence intervals based upon normal approximations may lead to estimates which are either negative or exceed one hundred. This may appear less important with confidence intervals for differences between proportions, though since these are dependent upon the confidence intervals for the response rates themselves they may also be distorted.

As an example of calculation and interpretation of odds ratios, consider the following. For two treatments in a clinical trial (in fact, the responses obtained in centre 1 in the example in Appendix 3), success rates were 26/30 (treatment A) and 21/30 (treatment B). For patients randomized to treatment A the probability of success is estimated at 26/30 (87%); the probability of failure as 4/30 (13%). The odds on a patient randomized to this treatment having a successful outcome are then given by the ratio of these probabilities, namely (26/30) / (4/30), = 26/4 = 6.5. For treatment B the odds on success are less at 21/9 or 2.3. The ratio of these odds (6.5/2.3), which is equal to 2.8, termed the odds ratio, indicates that the odds on a patient randomized to treatment A having a successful outcome are nearly three times those of a patient randomized to treatment B. If the two randomized groups had identical outcomes (as in centre 8), then the odds ratio would be equal to 1. Thus odds ratios with 95% confidence limit close to 1 indicate equivalent treatments. By contrast when the intervals are wide there may be insufficient evidence to support such a conclusion. Note also that when the 95% confidence interval for the odds ratio excludes the value of 1 then there is a statistically significant association (P < 0.05) by a conventional $\chi^2$ test.

Despite very widespread use in summarizing results from both clinical trials and epidemiological studies, the odds ratio has been criticized recently as not being helpful in clinical decision-making. As a result, an alternative measure, the ‘number needed to treat’, is now being advocated as a clinically useful measure of treatment effect, especially by the proponents of evidence-based medicine. This measure is discussed by Cook & Sackett.

### Appendix 3: Example data

Some of the recommendations for data analysis and presentation in this paper are illustrated by examples
based on the (fictitious) data summarized in Table A. These are intended to depict clinical outcomes in a typical randomized, multicentre clinical trial comparing the efficacy of two antibacterial agents.

The data were constructed to illustrate several characteristic features: (i) a moderate difference of 7% in overall success rates between the two drugs, a difference which may or may not be of clinical importance; (ii) a total sample size of 600 patients which is typical of many clinical trials of antibacterial agents; (iii) the participation of 10 centres with substantial variation in the numbers of patients recruited at each (between 20 and 140), as well as in the success rates achieved with the two drugs (standard: 40–97%, new: 70–98%) and the difference between them (new-standard: −17 to 40%). For convenience the numbers of patients allocated to each drug within each centre are balanced (this would be rare in practice despite stratified randomization within centre). To illustrate problems created by very small numbers of patients, which lead to zero and 100% success rates, patients in centre 9 have been re-distributed among three other centres (labelled 9a, 9b and 9c) in some of the examples.

<table>
<thead>
<tr>
<th>Centre</th>
<th>Standard treatment</th>
<th>New treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>success</td>
<td>failure</td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>9a</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>9c</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>240</td>
<td>60</td>
</tr>
</tbody>
</table>