Study design, methodology and statistical analyses in the clinical development of sparfloxacin

E. Genevois*, V. Lelouer*, Jean-Baptiste Vercken* and R. Caillon*

*Rhone-Poulenc Rorer Central Research, Biostatistics Department, 500 Arcola Road, Collegeville PA 19426, USA; *Rhone D. P. C. Europe, 20 Av R. Aron, 92165, Antony Cedex, France

Many publications in the past 10 years have emphasised the difficulties of evaluating anti-infective drugs and the need for well-designed clinical trials in this therapeutic field. The clinical development of sparfloxacin in Europe, involving more than 4000 patients in ten countries, provided the opportunity to implement a methodology for evaluation and statistical analyses which would take into account actual requirements and past insufficiencies. This methodology focused on a rigorous and accurate patient classification for evaluability, subgroups of particular interest, efficacy assessment based on automation (algorithm) and individual case review by expert panel committees. In addition, the statistical analyses did not use significance testing but rather confidence intervals to determine whether sparfloxacin was therapeutically equivalent to the reference comparator antibacterial agents.

Introduction

The accuracy and reliability of conclusions derived from the statistical analysis of a clinical trial depend primarily on the quality of the study design and reduction of potential sources of variability and bias. Variability and bias are usually introduced by having too vague or too wide a definition of the targeted study population, ambiguous exclusion of patients from analyses, subjective efficacy criteria (Chow & Yu, 1989) containing either an "observer compound" or a "patient compound", and inappropriate statistical methods (Pocock, Hughes & Lee, 1987).

A review of 21 community-acquired pneumonia clinical trials revealed a wide range of defined patient populations, as well as numerous efficacy assessment criteria, such as clinical, clinical and bacteriological, clinical and radiological, or simply bacteriological (Vercken, 1991). Therefore, there is increasing pressure from Regulatory Agencies for the study sponsor clearly to pre-determine the primary efficacy and safety criteria as well as the analytical methods.

In the USA, the Food and Drugs Administration (FDA) published guidelines for the evaluation of anti-infective drugs in 1977, which included recommendations for design, inclusion criteria, efficacy assessment criteria and specific methodological issues (FDA, 1977). Updated guidelines in the USA were published in 1992, prepared under a contract between the Infectious Diseases Society of America and the FDA (Beam, Gilbert & Kunin, 1992). European guidelines were published in 1993 (Beam, Gilbert
& Kunin, 1993) with the intention of maintaining consistency between US guidelines and modified European guidelines.

The following strategy was implemented during the clinical development of sparfloxacin in an attempt to avoid the classical pitfalls of anti-infective trials and to fulfill the regulatory requirements of various countries. These methodological principles were to apply to any of the target indications for sparfloxacin: acute community-acquired pneumonia, acute exacerbation of chronic obstructive pulmonary disease (COPD), acute purulent sinusitis, gonococcal and non-gonococcal urethritis and acute complicated urinary tract infection.

**Study designs**

All trials, regardless of the indication under study, were controlled, double-blind, parallel-group, randomised, multinational and multicentre. The purpose of each clinical trial was to evaluate the efficacy and safety profile of sparfloxacin in well-defined target infectious diseases rather than a mixture of dissimilar diseases such as acute and chronic bronchitis or bronchopneumonia and pneumonia. For example, in order to be included in a community-acquired pneumonia study, it was mandatory for patients to have an opacity on the chest X-ray.

Because of time and seasonal constraints, a large number of investigational centres were involved, with each one recruiting a small number of patients. Therefore, the randomisation schedule did not include any stratification factor such as disease severity. It was believed that the patient populations were defined with sufficient precision that this would not be a major concern.

Patients were screened and randomised when they presented with the disease under study. Baseline evaluations were then performed. Because antibiotic therapy should be started promptly in most acute infectious diseases, the treatment was sometimes initiated before bacteriological and/or radiological data were available, as was the case in the acute sinusitis study. Patients were evaluated 2 to 4 days after treatment initiation and again at the end of treatment, even when treatment was discontinued prematurely. An additional long-term follow-up visit was performed 2 to 8 weeks after treatment ceased, depending upon the disease, to assess whether patients had relapsed or not.

**Steering Committees**

One or two appropriate academic specialists were selected from each country to constitute the Scientific Review Committee of each trial. The use of these specialist committees, which are independent of the sponsor, is now widely recommended in clinical research and permit a clinical development programme to take into account the variations in medical practices in different countries. The Review Committee met four times to discuss the key points of each study corresponding to the main steps of the clinical trial design and analysis.

At the first meeting the design of the trial was discussed and finalised. The inclusion/exclusion criteria and the method and timing of collection for efficacy and safety data were defined. The selection of an appropriate comparator antibacterial and dosage was particularly difficult on account of the various European countries participating in the programme.
At the second meeting, the criteria for assigning patients to a pre-defined category either for evaluability (evaluable or not and reason for unevaluability) or to a specific subgroup of interest for which the definition was not straightforward (Anthonisen type I patients (Anthonisen et al., 1987) in the COPD study for example) were elaborated with the sponsor.

The definitions of efficacy outcome as success or non-success (binary response), taking into account the clinical and/or radiological and/or bacteriological responses as well as the patient’s underlying conditions, were agreed with the sponsor. Possible protocol violations such as treatment discontinuation and prohibited medications which could interfere with the efficacy assessment were also pre-defined.

At the third meeting, all the ‘complex cases’ were reviewed in a blinded fashion, once the sponsor’s data base was finalised. These ‘complex cases’ were those which were not automatically classifiable from the algorithms in the pre-determined categories for evaluability and efficacy outcome.

In the acute sinusitis study, the Review Committee participated in reading all sinus X-rays taken during the study, either at baseline or at the end of treatment period, in order to provide a homogeneous assessment of patients. The Review Committee was also involved in reviewing the results of serological and bacteriological tests in the community-acquired pneumonia studies.

Although this type of centralised assessment is methodologically optimal for eliminating the observer effect and thus reducing variability, this was not feasible for all trials for obvious logistical reasons.

After the study was unblinded and statistical analyses were completed, the results were presented and discussed at a fourth meeting.

A central assessment of all clinical results was performed by the Review Committee. This ensured that all patients were evaluated consistently, regardless of which country the study was conducted in.

**Patient population**

It is well known that in any double-blind, randomised clinical trial, the primary efficacy analysis should be performed on the intent-to-treat population. This is the best way to prevent bias from being introduced in a post-hoc selection of patients and to keep the statistical test as powerful as initially planned. This is one of the main requirements of most Regulatory Agencies and is specified in the French, European and American guidelines (Beam et al., 1992, 1993). However, there is an obvious need to evaluate the effects of the drug in a pre-determined subset of ‘evaluable patients’ who actually present with the disease under study, comply with their treatment and are evaluated clinically and/or radiologically after the end of the treatment to determine the response. These ‘evaluable patients’ are fully representative of the disease under study and generally fulfill both the inclusion criteria and most of the pre-defined protocol procedures.

The results of the analysis of evaluable patients were to be as consistent as possible with those of the intent-to-treat analysis. In the case of discrepancy, a careful analysis of the reasons for exclusion in each treatment group was to be performed to detect the potential source of bias.
Rules for evaluability and specific subgroups at baseline

The intent-to-treat population is usually defined as 'all randomised patients'. For the purposes of the sparflloxacin clinical trials programme, it was defined as 'all patients who received at least one dose of the study drug', thus excluding from all analyses those few patients who were randomised but did not receive the treatment. The number of patients excluded from the intent-to-treat analysis was similar in each treatment arm and very low, representing <1% (0 to 0.98%) of the entire population included. The impact of these exclusions was negligible on the results.

The reasons for not including a patient in the subset of 'evaluable population for efficacy' depended upon the protocol procedures and the inclusion/exclusion criteria. These may be numerous. However, many protocol criteria aim at selecting a more representative patient population, for clinical or ethical reasons. This does not mean that the patient would definitely be excluded from analysis, e.g. if a patient was aged 17 years old rather than meeting the minimum age of 18 years of age. The primary reasons for exclusion from the evaluable population in descending order of importance: inappropriate diagnosis at baseline, e.g. suspected pulmonary embolism instead of pneumonia; major protocol violations as defined by the Review Committee: prohibited medication, incorrect date of visit, non-compliance with the treatment schedule; partial or total unavailability of the efficacy data used to derive the overall response, e.g. temperature not documented in the case of pneumonia or dyspnœa not recorded in the case of exacerbation of COPD.

If there was more than one reason that a patient was excluded from the 'evaluable population', a primary reason was assigned so that these patients could be tracked, comparisons could be made between treatment groups and, in the case of discrepancies between the results for the intent-to-treat and evaluable populations, an appropriate explanation could be found.

The mean rate of exclusion from the 'evaluable population' analysis across all studies was 26.8% (51.8 to 13.1% of intent-to-treat patients). This high exclusion rate was primarily caused by the stringent inclusion criteria in the acute exacerbation of COPD study in which patients had to present with a documented obstructive syndrome at baseline.

Before unblinding, the Review Committee assessed some of the patients for whom the diagnosis of the disease under study was questionable. Indeed, at the time the patient was treated, the results of bacteriological cultures and X-ray could still be unknown. Therefore, a post-hoc classification for inclusion diagnosis was sometimes necessary. It was also determined whether sufficient evidence of treatment non-success could be derived from incomplete efficacy data. Consequently, evaluability and efficacy classifications were highly correlated. For example, in the situation where a patient with pneumonia lacked complete clinical data at the end of treatment but still had a fever, the treatment was deemed a non-success and the patient was included in the evaluable population. If the same patient had no fever, however, clinical success could not be confirmed and the patient was not included in the subset of evaluable patients.

The algorithms developed for assessment of both evaluability and efficacy were sometimes complex with regards to how to handle missing information but this dramatically reduced the number of patients who had to be assessed by the Review Committee.

In the interim, another set of criteria was used to assess the pathogenicity of each
isolated organism. The algorithm developed incorporated each possible method of sampling and the criteria reflecting the quality of the sample, e.g. colony count, epithelial cells and polymorphonuclear counts. Where there were multiple sources of samples, these were ordered according to the reliability of the sample technique. The goal was to distinguish between non-bacteriologically documented and bacteriologically documented infections according to strict microbiological criteria. The stringency of these criteria probably underestimated the number of patients with documented infections.

**Binary primary efficacy criterion**

It has been a common approach in previous anti-infective clinical trials to evaluate the clinical response of patients as cured, improved or failed. However, satisfactory response rate generally comprised the categories of both cured and improved. In the evaluation of efficacy of sparfloxacin, it was felt that in order to reflect usual medical practice, a simple binary response would be determined for each patient. The outcome of the antibiotic therapy was classified as success or non-success in all patients, at both endpoints (end of treatment and follow-up) for all sparfloxacin clinical trials. Patients who were classified as non-success at the end of treatment were automatically classified as non-success at follow-up. On the other hand, a patient could be classified as successful cure at the end of treatment and unsuccessful at follow-up in the case of relapse or reinfection. However, in the study of acute exacerbation of COPD, the classical evaluation of patients as cured, improved or failed was maintained because it is an accurate reflection of the effect of antibiotic therapy on the evolution of this disease.

The major concern raised by using a binary outcome method was in the case of patients classified as success at the end of treatment who did not return for follow-up (although every effort was made to obtain information about their clinical status). In the intent-to-treat analysis, these patients were automatically classified in the non-success category at follow-up. The worst case scenario (all comparator patients classified as success, all sparfloxacin patients classified as non-success) was not used because this method appeared to be inappropriate for equivalence trials, especially with the very small differences observed between antibiotic therapies and the 5 to 10% incidence of patients cured at the end of treatment and then lost to follow-up.

**Case-by-case review by the Review Committee**

The set of guidelines developed for automatically classifying patients as success or non-success reflected clinical reality and allowed the classification of at least 80% of the patients in whom the diagnosis was certain (all inclusion criteria fulfilled) and response was satisfactory (e.g. disappearance of all clinical signs and symptoms, normal X-ray at follow-up if required, no requirement for additional antibacterial agent). The patient's profile was defined during the second Review Committee meeting (at the time, recruitment of patients was still ongoing) and the experts listed all information necessary for a conclusion to be made. All these parameters were displayed on a single sheet 'patient's profile' that the sponsor provided for the case-by-case review at the subsequent meeting. This third meeting could not be held until all the data entry, data
validation and data classification procedures were finalised within Rhône D.P.C. Europe Biometrics Department, before unblinding the treatment codes.

An average of 15% of the cases were reviewed during the third meeting (about 100 complex cases per study), requiring a consistent approach in reviewing all the patients. The experts had to reach a consensus on each patient and there could not be any contradiction between the individual patient classification and the rules used to automatically classify the majority of patients.

Statistical analysis and interpretation of results

The statistical methods used in previous anti-infective trials have been criticised (Pocock et al., 1987) for a wide use of significance testings, with multiple examinations of the data precluding a control of type I error, no investigation of type II error and other deficiencies. It seems, however, that there is a trend towards using a common method to analyse such data, one of which is the equivalence approach. At present, the expected response rate with antibiotic therapy in the treatment of respiratory tract infections is approximately 85%. Demonstrating statistically that a new drug is superior would require very large cohorts of patients which is unrealistic considering the seasonal recruitment of patients, the necessity of performing two clinical trials in the main indications and cost. Moreover, where placebo controlled studies are ethically contraindicated it is very unlikely that a new drug can be proven to be 5 or 10% more efficient than the current reference therapy in separate clinical trials.

Many published clinical trials have reported statistical results concluding that two antibacterial drugs are of equivalent efficacy since no significant difference was detected. This point is not statistically valid and in clinical trials designed to demonstrate superiority, failure to achieve this goal does not justify a conclusion that the treatments are of equivalent efficacy.

This has led biostatisticians to develop new statistical methods with which to establish that two treatments have equivalent efficacy. In fact, the goal statistically is to demonstrate that the new drug is at least no worse than the reference drug by more than a specified percentage, this being the maximal acceptable difference between the reference drug and the new agent. In the sparfloxacin clinical development plan, a difference of 10% was considered acceptable. In addition, it was important to determine whether the reference antibiotic therapy was or was not superior to sparfloxacin by 10% or more. In this way, in each separate clinical trial, the eventual superiority of sparfloxacin by 10% or more was not investigated (unilateral test).

This method for calculating sample sizes was simple and relied upon confidence intervals at 90% (90% CI) of the observed difference between study drugs (Makuch & Simon, 1978). The interpretation of the results is derived directly from the 90% CI; assuming a maximal acceptable difference of 10%, it was concluded that the reference antibiotic therapy was not superior to sparfloxacin by 10% or more when the upper limit of the 90% CI of the difference was < 10%. On the other hand, the hypothesis that the reference drug was superior by 10% or more could not be rejected when 10% was actually within the confidence interval. The use of 90% CI instead of the more commonly used 95% confidence interval was a reflection of the ‘one-sided’ nature of the analysis. By focusing the analysis in this way, the type I error was only 5% and represented the probability to reject the null hypothesis of superiority of the reference drug. The main advantage of this statistical approach was that it allowed a conclusion
in all cases and an estimation of the magnitude of the difference, using the width of the confidence interval. For example, with 90% CI \((-27; 34\%\)) the treatments would not be equivalent at 10%.

When equivalence could not be demonstrated (10% within the upper limit of the 90% CI), the probability of accepting non-equivalence when it was false was calculated (power). The calculation of the 90% CI and subsequent power was applied to the primary efficacy criteria (success or non-success) in all sparfloxacin trials, for both the intent-to-treat and evaluable populations, at both endpoints (end of treatment and follow-up). Therefore, four analyses were presented for each trial. The two intent-to-treated analyses provided the most acceptable conclusions from a methodological standpoint but underestimated the success rates because of the stringent patient classification. The evaluable population analysis provided a more accurate estimate of the success rates of treatments but also allowed for corroboration of the results of the intent-to-treat analysis. When an inconsistency between the analyses was identified, all subcategories of patients in the intent-to-treat analysis and not in the evaluable population were explored to determine whether one specific subgroup influenced the comparison. It was striking to observe that none of the trials showed inexplicable discrepancies.

It was also possible to perform a pooled data analysis of homogeneous studies in order to assess the efficacy of sparfloxacin versus a set of reference antibiotics. A pooled data analysis of the two community-acquired pneumonia studies was possible because all of the following requirements were fulfilled: same definition of community-acquired pneumonia for both studies, similar mean duration of treatment and same methodology for analysis (e.g. same definition of evaluable populations, same endpoints and same criteria for evaluation).

The statistical methodology for this pooled data analysis was based on the calculation of the Odds ratio (sparfloxacin success/sparfloxacin non-success over pooled reference drugs success/pooled reference drugs non-success) and the calculation of the 90% CI for the Odds ratio. An Odds ratio \(>1\) indicated a trend toward superior efficacy of sparfloxacin, confirmed when the lower limit of the 90% CI was \(>1\).

**Conclusion**

The purpose of the clinical development programme of sparfloxacin was to evaluate whether this new aminofluoroquinolone is effective and well tolerated in the treatment of specific community-acquired infectious diseases, particularly in the field of respiratory tract infections. By adhering to, and occasionally anticipating, the guidelines for the clinical evaluation of anti-infective drugs, with the advice of experts who participated to the advisory boards and Review committees, this goal was reached. The methods of statistical analyses used for the clinical trials allowed for a clear and convincing conclusion as to whether sparfloxacin was at least equivalent to reference antibacterial agents in the conditions studied.

**References**


