J Pharmacokinet Pharmacodyn (2011) 38:595–612
DOI 10.1007/s10928-011-9210-8

Modelling of the outcome of non-inferiority trials by integration of historical data

Alberto Russu · Erik van Zwet · Giuseppe De Nicolao · Oscar Della Pasqua

Received: 3 April 2011 / Accepted: 23 July 2011 / Published online: 21 August 2011

Abstract The approval and differentiation of new compounds in clinical development often demands non-inferiority trials, in which the test drug is compared against a reference treatment. However, non-inferiority trials impose major operational burden with serious ethical and scientific implications for the development of new medicines. Traditional approaches make limited use of historical information on placebo and neglect inter-trial variability, relying on the *constancy assumption* that the control-to-placebo effect size is maintained across trials. We propose a model-based approach that overcomes such limitations and may be used as a tool to explore differentiation during clinical development. Parameter distributions are introduced which reflect the heterogeneity of trials. The method is illustrated using data from impetigo trials. Based on simulation scenarios, this Bayesian technique yields a definitive, consistent increase in the statistical power over two accepted statistical methods, allowing lower sample size requirements for the assessment of non-inferiority.

Keywords Non-inferiority · Bayesian mixed-effects model · Trial variability · Placebo effect · Historical information · Model-based drug development

A. Russu · G. De Nicolao
Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy

E. van Zwet
Bioinformatics Center of Expertise, LUMC, Leiden, The Netherlands

O. Della Pasqua
Clinical Pharmacology and Discovery Medicine, GlaxoSmithKline, Stockley Park, UK

O. Della Pasqua (✉)
Division of Pharmacology, Leiden/Amsterdam Center for Drug Research,
PO Box 9502, 2300 RA Leiden, The Netherlands
e-mail: odp72514@gsk.com

Introduction

Medical and health care policies often demand drugs to be effective, rather than just efficacious. However, there are circumstances in which the assessment of effectiveness implies demonstration of non-inferiority rather than superiority. These considerations are relevant to product differentiation when the experimental compound is not expected to outperform the established standard of care based on the primary protocol endpoint [1]. Instead, the treatment is meant to offer ancillary advantages in terms of its clinical pharmacology profile, safety, tolerability, cost, or convenience. In this situation, demonstrating a statistically significant effect size imposes huge sample sizes, as the difference in efficacy between the two treatments is small [2].

Non-inferiority has been the subject of late drug development in statistical and clinical trial publications. However, it has clear implications for early clinical studies when standard of care is effective and add-on or combination therapy is to be used as primary indication. The concept of non-inferiority trials was introduced to compare a so-called golden standard against the new drug, excluding the requirement for comparison against a placebo arm [3, 4]. More recently, regulatory guidance has been issued on the use of concurrent control to support the conclusion that the new test drug is also effective [5, 6]. This requirement entails operational challenges, including financial and ethical considerations, with direct implications for the design of clinical trials aimed at the differentiation of novel compounds. Despite the justification for the use of an active comparator, the availability of a placebo arm still constitutes important evidence of drug response and, of course, enables establishing assay sensitivity as well as the underlying exposure–response relationship.

Given that non-inferiority trials often do not include placebo, uncertainty exists about the true magnitude of the effect of each individual treatment. This represents a major challenge for decision making in early clinical development, including further understanding of the dose-exposure–response curve in the population of interest. In fact, based on current practice, one has partial evidence about the overall distribution of the effect size by the time drugs reach the market and only a few clinical trials have been performed. In addition, it has been established that patient population and other factors, including placebo effect also contribute to variability in treatment response. The consequence of this uncertainty is an inflation of the size of such trials, which other than just being expensive raise the ethical question about the need to enrol so many subjects to reach the level of evidence required for non-inferiority. Moreover, these hurdles prevent a meaningful evaluation of the therapeutic value of the compound prior to Phase III.

From a methodological perspective, non-inferiority trials [7] require the new compound to be statistically non-inferior relative to the active comparator (e.g. at least 90% efficacy). Additionally, superiority to placebo (of at least 50% in terms of clinical efficacy, for example) [5, 6] may be required for regulatory purposes in therapeutic areas such as psychiatry. In the literature related to non-inferiority trials with binary outcome, which is further considered in this manuscript, the assessment of non-inferiority involves test statistics such as success proportion or odds ratio for

the different treatment arms. Efficacy endpoints may be the proportion of cured patients or the reduction in the rate of adverse events or mortality after a predefined treatment period. Comparison between treatment arms is then performed using statistical hypothesis testing [8]. Utilisation of historical information is limited or absent: previous trials may be simply used to work out an estimate of the placebo effect in the current non-inferiority trial, should it contain no placebo arm. This is also known as *putative placebo* analysis [9, 10]. Furthermore, traditional methods do not address the issue of inter-trial variability. On the contrary, they rely on the so-called *constancy assumption*, that is, that the historical difference between control and placebo treatment arms is also maintained in the current trial [8, 11]. Such effect size is in turn used to estimate the non-inferiority margin for the comparison between treatment arms [8]. Moreover, testing multiple hypotheses (non-inferiority to the active control and superiority to placebo) is made difficult by the frequentist nature of such analyses. If fact, the reliance on *P*-values has been widely criticised for comparison purposes [12].

To date, most of the debate concerning the differentiation of treatment effect in clinical development revolves around the question whether proving non-inferiority is feasible or ethical. Instead, we feel that attention should focus on how to properly analyse trial data, exploit the available knowledge (also from past trials), obtain realistic estimates of required sample size to plan a non-inferiority trial and most importantly demonstrate whether differentiation is achievable at early stages of development. The aim of our investigation is therefore to (i) develop a general Bayesian framework for non-inferiority (and possibly superiority) assessment, (ii) show its ability to incorporate data from historical trials, and (iii) compare the performances of the proposed approach to two accepted statistical methods, in terms of required sample size and power.

A Bayesian mixed-effects model for binary endpoints is proposed that overcomes the limitations of traditional methods. Test statistics are derived through logistic regression. The constancy assumption is relaxed by introducing parameter distributions, which reflect the heterogeneity of available trials. Of note, the Bayesian approach adopted here enables the estimation of the probability that the experimental drug is non-inferior to the active control while being superior to placebo, through the joint posterior distribution obtained from the new and previous trials. Moreover, it is possible to account for the uncertainty in the estimation of the inter-trial variability, which is not directly feasible in a non-Bayesian or frequentist context. A comprehensive assessment of uncertainty is a key factor in the evaluation of non-inferiority. The possibility to make well-defined probability statements without resorting to null hypotheses and *P*-values is therefore a further advantage.

These concepts are illustrated by the analysis of non-inferiority of two treatment options for impetigo. Impetigo is a staphylococcal skin infection that is particularly frequent in children for which several treatment options are available, including topical creams and oral antibiotics [13]. Based on a recent review, it was possible to implement the proposed method in conjunction with a retrospective analysis of 13 clinical trials [14] and two trials involving an investigational compound. Subsequently, results from simulations were compared to two accepted statistical

methods, namely a non-hierarchical Bayesian technique [15] and an implementation of the traditional frequentist method based on difference of sample proportions with putative placebo analysis [10, 11].

Methods

Patients and data

Data from two Phase III trials of an experimental drug, as well as data from 13 trials reported in a review by Koning et al. [14] were analysed (Table 1). The clinical endpoint was success or failure after 5 days of treatment. The trials from literature involved either placebo, fusidic acid or both. The other two studies were a randomised, double-blind trial of the test drug vs. placebo (study A) and a randomised, observer-blind trial of the test drug vs. fusidic acid (study B). Further details can be found in the GlaxoSmithKline clinical trial register (http://www.gsk-clinicalstudyregister.com/files/pdf/21118.pdf and http://www.gsk-clinicalstudyregister.com/files/pdf/21117.pdf). Overall, 18 treatment arms were considered (6 placebo, 10 fusidic acid, 2 test drug). In each trial, the primary outcome was the proportion of clinical cure at a predefined study day, reported as the observed number of clinical successes divided by the arm size in Table 1.

Bayesian population model of available trials

A Bayesian model was developed that does take into account historical trials but without resorting to the constancy assumption. Here, "population" refers to the collection of available trials, not patients. Probabilities of clinical success were obtained for each trial of the population using logistic regression. All trials are seen as if we were observing "noisy" versions of a true, *typical trial*. More precisely, within each trial, probabilities for placebo, comparator and test drug are modelled as:

$$\begin{cases} logit(\pi_p^{(k)}) = \alpha + \tau_k \\ logit(\pi_c^{(k)}) = \alpha + \beta + \tau_k \\ logit(\pi_t^{(k)}) = \alpha + \gamma + \tau_k \end{cases} \quad (1)$$

where $k = 1 \ldots K$; k represents each individual trial and K the total number of trials. The subscripts p, c and t indicate placebo, comparator and test compound, respectively.

Success probabilities in actual trials are simply obtained by inverting the *logit* relationship. The characterisation of the typical trial is obtained by removing the random effects τ_k from Eq. 1. Therefore, the fixed-effects α, β and γ, characterising the success probabilities of the typical trial, are termed *typical parameters*. The *typical probabilities* are easily obtained as:

$$\pi_p = \frac{e^{\alpha}}{1+e^{\alpha}} \quad \pi_c = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} \quad \pi_t = \frac{e^{\alpha+\gamma}}{1+e^{\alpha+\gamma}} \quad (2)$$

in which the superscript (k) has been dropped.

Table 1 Synopsis of experimental impetigo trials. Historical trial details can be found in the review by Koning et al. [14]

No.	Study	Placebo	Fusidic acid	NCE
1	Christensen (1994) [16]	–	2 to 3 td, evaluation time not reported 105/128	–
2	Eells (1986) [17]	3 td, day 8 8/19	–	–
3	Gilbert (1989) [18]	–	3 td, day 7 6/11	–
4	Gould (1984) [19]	1 td, evaluation time not reported 7/21	–	–
5	Koning (2002) [20]	3 td + povidone-iodine 2 td), day 7 10/80	3 td + povidone-iodine 2 td, day 7 42/76	–
6	Moraes Barbosa (1986) [21]	–	3 td, day 7 10/12	–
7	Morley (1988) [22]	–	3 td, day 6 to 8 45/51	–
8	Park (1993) [23]	–	20 to 40 mg/kg/d, day 7 11/18	–
9	Rojas (1985) [24]	3 td, day 7 to 12 15/52	–	–
10	Ruby (1973) [25]	3 td, day 5 0/20	–	–
11	Sutton (1992) [26]	–	3 td, day 8 90/93	–
12	Vainer (1986) [27]	–	Dosing not reported, day 7 26/43	–
13	White (1989) [28]	–	3 td, day 7 33/49	–
14	GSK A	2 td, day 7 37/73	–	2 td, day 7 119/140
15	GSK B	–	3 td, day 9 150/156	2 td, day 7 315/320

NCE new chemical entity, *td* times per day, *day* the evaluation time of the clinical endpoint

Cure rates are reported as observed number of clinical successes divided by the arm size

We assigned non-informative normal priors to fixed-effects α, β and γ, so as not to bias estimates of the typical probabilities. Random trial effect τ_k (the "noise") had a normal prior with zero mean and variance ω^2. Note that, although different distributions could be used for the parameters of the *logit* relationship, the normal

distribution is an obvious choice (see for example [15]). In order to account for uncertainty in the inter-trial variability, the parameter ω^{-1} (inter-trial standard deviation) was assigned a uniform prior distribution, as suggested in [29]. The range of the uniform density was set wide enough to be uninformative while avoiding unreasonably high values of ω^{-1}. This was achieved by setting the lower bound to zero and the upper bound to ten times the value of a preliminary restricted maximum likelihood estimate of ω^{-1}. See [29, 30] for an extensive discussion on the choice of prior distributions for variance parameters in hierarchical models.

Non-inferiority and effect retention test

Differently from current practice, our purpose is not only to assess non-inferiority of the experimental drug with respect to the active control, but also to quantify superiority of the new drug with respect to placebo in terms of *effect retention* (the fraction of control-to-placebo effect size retained by the new compound). Therefore, we define the following test statistics:

$$\begin{aligned} T_1 &= \pi_t - \mu_1 \pi_c \\ T_2 &= (\pi_t - \pi_p) - \mu_2(\pi_c - \pi_p) \end{aligned} \tag{3}$$

where μ_1 is termed non-inferiority margin, whereas μ_2 is the preservation factor (i.e. fraction of the control-to-placebo effect size retained by the test drug) [8]. In general, effect fractions μ_1 and μ_2 range between 0 and 1.

Although recommendations from the International Conference on Harmonization and the European Medicines Agency suggest that the non-inferiority margin μ_1 should be justified on both statistical and clinical grounds [5, 6, 31], its proper choice is still matter of debate [32, 33]. Widely used values have been reported to be 80 or 90% for μ_1 [11, 15], and 50% for μ_2 [11].

In order to conclude non-inferiority *and* effect retention, we calculate the probability that both T_1 and T_2 are greater than zero, given the current data. This is readily achieved using the joint posterior density of T_1 and T_2:

$$P(T_1 > 0, T_2 > 0|\text{data}) = \int_0^\infty \int_0^\infty f(T_1, T_2|\text{data}) dT_1 dT_2 \tag{4}$$

Such probability is then compared to the threshold value P_{cutoff} in order to accept or reject the experimental drug. Throughout this investigation, we have used 90 and 50% for μ_1 and μ_2 respectively, and 95% for P_{cutoff}.

Prospective use in clinical trials: statistical analysis

It is of interest to evaluate this approach prospectively, by means of simulation scenarios. Simulations also enable the assessment of properties such as specificity and sensitivity, as well as estimation of the required sample size for a new trial. We compared our model to a non-hierarchical Bayesian [15] and a standard implementation method based on the comparison of sample proportions of clinical success, with putative placebo analysis [10, 11].

The following methods were tested to assess Type I error and power in the simulated benchmark:

1. The Bayesian population method (Eqs. 1–4).
2. A Bayesian, non-hierarchical method proposed by Simon [15]. The method was originally developed for binary outcome data, which approximates log odds of failure by the assumption of a normal distribution. We centred the prior of parameter β (representing the comparator-to-placebo effect size) around the true effect size used in simulation, and assigned a small standard deviation (33% relative to the prior mean). Parameters α (placebo effect size) and γ (test drug-to-placebo effect size) were assigned uncorrelated, non-informative normal priors. Acceptance of the test drug was checked by evaluating T_1 and T_2 using probabilities from the model, and comparing $P(T_1 > 0, T_2 > 0 \mid \text{data})$ (Eq. 4) with P_{cutoff}.
3. A traditional analysis ("standard method") based on the comparison of sample proportions of clinical success, with putative placebo [10, 11]. Test statistics T_1 and T_2 were calculated as point estimates using Eq. 3: success probabilities π_c and π_t were estimated from data of the non-inferiority trial, whereas $(\pi_c - \pi_p)$ was set to the true comparator-to-placebo effect size used to simulate data. Estimates of T_1 and T_2 ($\hat{T}_1$ and $\hat{T}_2$ respectively) were compared against their null distribution H_0, i.e. the distribution of estimates under the null hypothesis that $T_1 = 0$ and $T_2 = 0$. The test drug was then accepted if

$$p = \int_{-\infty}^{\hat{T}_2} \int_{-\infty}^{\hat{T}_1} f(T_1, T_2 | H_0) dT_1 dT_2 \tag{5}$$

was greater than P_{cutoff}. Note that, because of the different nature of the standard analysis with respect to the two Bayesian alternatives, the acceptance test has to be performed through a P-value (Eq. 5) rather than a posterior probability (Eq. 4).

Table 2 reports the values of typical probabilities for the analysis of Type I error and power. Although both the non-hierarchical Bayesian and the standard method account for success probabilities only in the non-inferiority trial, in the former approach historical information is used to elicit prior distributions, whereas in the latter it serves to perform the putative placebo analysis, by means of the constancy assumption.

Simulation study: design scenarios

We have built a benchmark by simulating different real-life scenarios. Four design features were considered: sample size of the non-inferiority trial (5 choices), number of available historical trials (6 choices), sample size of historical trials (2 choices) and inter-trial variability (2 choices). By combining such features, we obtained $5 \times 6 \times 2 \times 2 = 120$ scenarios. For each of them, 1,000 simulations were run. Table 2 reports design values used to simulate datasets. Values of effect

Table 2 Synopsis of the design characteristics for the analysis of Type I error and power

Analysis	Design characteristic	Value
Type I error	π_p	0.5
	π_c	0.625
	π_t	0.5625
Power	π_p	0.5
	π_c	0.9
	π_t	0.9
	μ_1	90%
	μ_2	50%
	P_{cutoff}	95%
	Sample size of non-inferiority trial	100
		200
		300
		400
		500
Both	Number of available historical trials	1
		2
		3
		6
		12
		18
	Sample size of historical trials	50
		200
	Inter-trial variability (ω^2)	0
		0.1

fractions μ_1 and μ_2 were 90 and 50% respectively, and 95% for P_{cutoff}. Given its hierarchical structure, we used the model to generate datasets. However, for fairness, we also analysed the case of no inter-trial variability. Design characteristics that were investigated are further summarized in Table 3. Scenarios were categorized into six groups and include:

a. *Sample size*: Non-inferiority trials were simulated using 100, 200, 300, 400 and 500 total patients, equally randomized between the two treatment arms (active comparator and experimental drug).
b. *Number of available historical trials*: Table 3 shows the alternatives that were considered. For case A, the only available historical trial featured a single placebo arm, whereas in case B a comparator-vs.-placebo trial was also simulated. In the remaining cases, we simulated placebo-only (*P*), comparator-only (*C*) and *C*-vs.-*P* trials in equal proportion.
c. *Sample size of historical trials*: Historical trials were simulated using 50 and 200 total patients in cases B to F, with equal randomization between the two treatment arms (placebo and active comparator). In case A, 25 and 100 patients were simulated for the placebo arm.

Table 3 Summary of available historical trials in the simulation benchmark

Label	Total trials	*P*-only trials	*C*-only trials	*C*-vs.-*P* trials
A	1	1	0	0
B	2	1	0	1
C	3	1	1	1
D	6	2	2	2
E	12	4	4	4
F	18	6	6	6

d. *Inter-trial variability*: We simulated historical trials with and without inter-trial variability. In the former case, a deviation of ±30% from a hypothetical placebo success probability of 0.5 (equivalent to $\omega^2 = 0.1$) was used. Similar results were obtained with different values of inter-trial variability. Note that, in general, the value of ω^2 cannot directly be interpreted as variability in probability, given the "warping" introduced by the *logit* operator.

Software implementation

Fitting of the Bayesian population model was performed using WinBUGS 1.4.3 [34], the Bayesian non-hierarchical method [15] and the standard analysis [10, 11] were implemented using R 2.8.0 [35]. In addition, R was used to calculate probability integrals (Eqs. 4, 5), as well as to obtain graphical summaries.

Results

Model evaluation

The Bayesian mixed-effects model was estimated from the 15 trials in Table 1. These trials involved either placebo, fusidic acid or both. Summaries of the posterior distributions are reported in Table 4 for typical parameters α, β, γ, typical probabilities π_p, π_c, π_t, inter-trial variability ω^2 and test statistics T_1 and T_2. Parameters were satisfactorily estimated despite the availability of only 1 or 2 arms for each trial. As expected, standard deviations of typical success probabilities are relatively small (24.57%, 7.28%, 8.13% of their respective posterior means), since all trials contribute to their estimation.

Model performance and the goodness-of-fit were obtained by plotting the observed proportion of clinical success against the effect distribution for each arm in each trial (Fig. 1). The estimated proportions were satisfactory, with observations lying well within the support of the posterior distributions. The width of posterior distributions reflects the different number of subjects in each arm: the larger the arm size the narrower its associated posterior distribution, meaning less uncertainty about the effect magnitude.

Furthermore, we assessed the consistency of model simulations with respect to the observed data by performing a predictive check, that is, by calculating the

Table 4 Summary of posterior distributions and non-inferiority analysis on the experimental impetigo trials

Quantity	Mean	SD	2.5%	Median	97.5%
Posterior summaries					
α	−0.480	0.423	−1.330	−0.450	0.219
β	1.875	0.312	1.300	1.873	2.473
γ	2.086	0.323	1.448	2.098	2.724
π_p	0.387	0.0952	0.209	0.389	0.555
π_c	0.795	0.0579	0.667	0.801	0.890
π_t	0.823	0.0669	0.665	0.830	0.925
ω^2	1.884	0.947	0.717	1.670	4.223
T_1	0.108	0.0543	−0.00869	0.111	0.203
T_2	0.232	0.0500	0.131	0.234	0.330
Non-inferiority analysis					
μ_1	90%				
μ_2	50%				
P_{cutoff}	95%				
$P(T_1 > 0, T_2 > 0 \mid \text{data})$	96.8%				

Observe that in a typical trial, the probability that the experimental drug is non-inferior to the active comparator while being superior with respect to placebo is 96.8%, i.e., greater than the acceptance cut-off of 95%

distribution of success proportions in trials simulated from the model parameters (also termed *predictive distributions*) (Fig. 1). In principle, data simulated from the model should agree with the observed data, if the parameter estimates are realistic. In view of such considerations, predictive checks differ conceptually from usual goodness-of-fit plots, and therefore represent an additional diagnostic tool. Note that the larger width of predictive distributions with respect to posterior distributions is not due to imprecision in parameter estimation, but to the fact that predictive distributions incorporate inter-trial variability.

Observe that, in a typical trial, the probability that the experimental drug is non-inferior to the active comparator while being superior with respect to placebo is 96.8%, i.e., greater than the acceptance cut-off of 95% (Table 4). Such posterior probability is obtained by evaluating Eq. 4 on the joint posterior density of T_1 and T_2 (Fig. 2). Based on these results, we would accept non-inferiority of the new compound.

We are well aware of the potential pitfalls of analyzing historical information, such as exchangeability and publication bias. A thorough coverage of such topics is beyond the scope of this work, as extensive literature on the topic exists (see for example [30, 36]). To illustrate our statistical procedure, we considered all trials as exchangeable. The interested reader is referred to [14], which details the characteristics of the analyzed impetigo trials. Differences among trials (e.g. due to different standards of care, patient population, dose, drug exposure, etc.), which

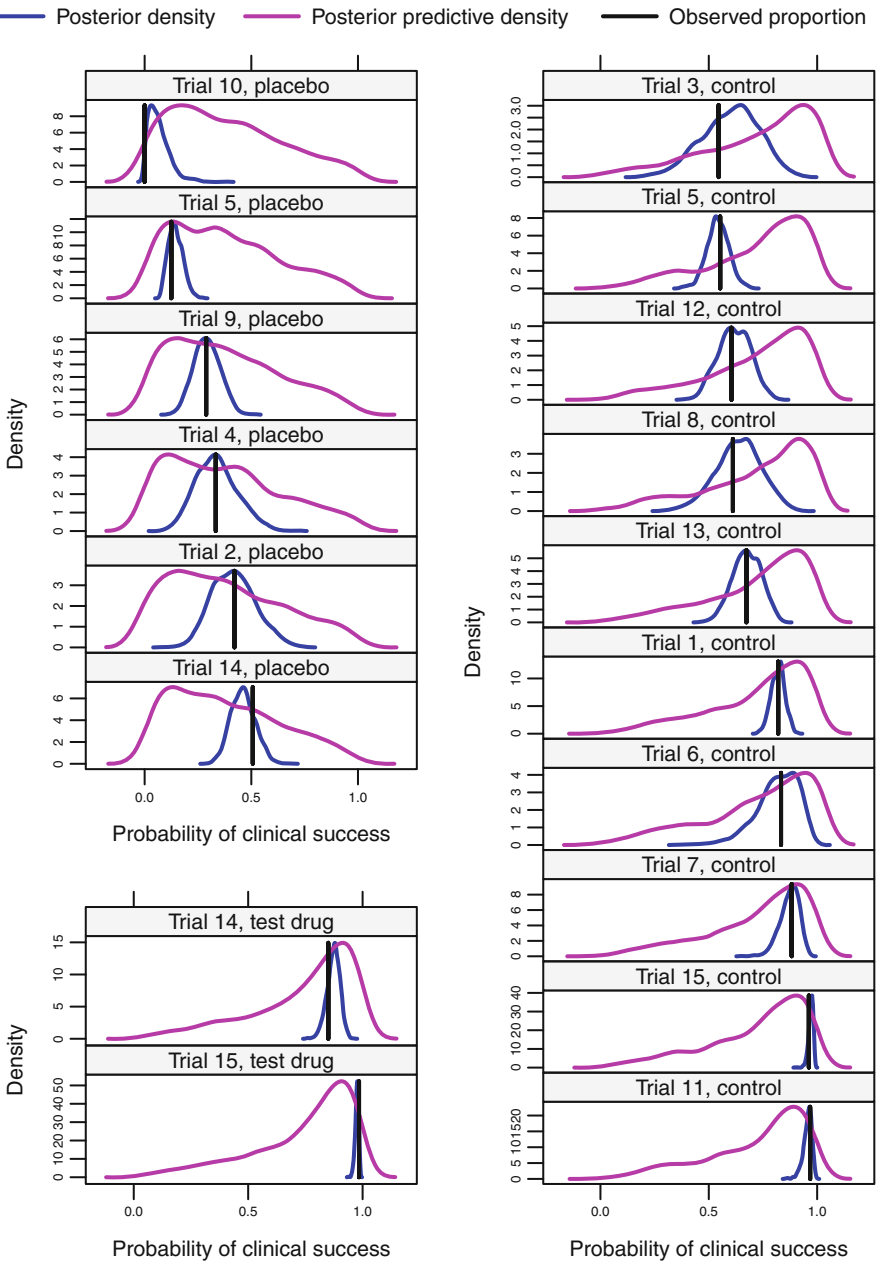

Fig. 1 Results obtained from fitting the Bayesian population model to the 15 experimental datasets. Placebo, control, and test drug are grouped together, clockwise. Each panel shows the posterior of the success probability, the posterior predictive density, and the observed success proportion. Within each treatment group, panels are ordered in increasing effect magnitude

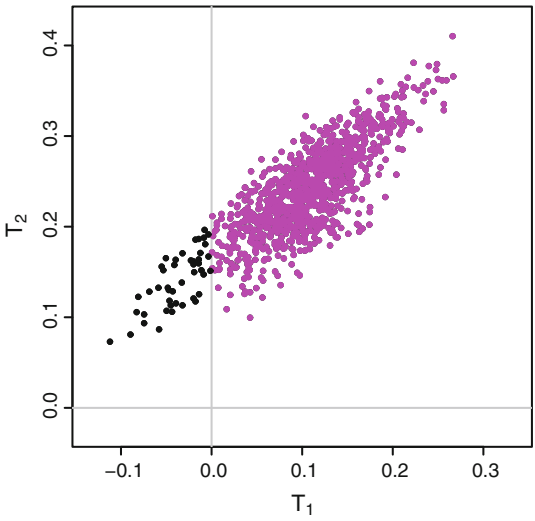

Fig. 2 Joint posterior density of the two test statistics T_1 and T_2 (Eq. 3), obtained by samples. The probability that both statistics are greater than zero (Eq. 4) is 96.8%. The associated region of the posterior density is represented in magenta (Color figure online)

are easily encountered when multiple historical studies are considered, are assumed to be captured by the inter-trial variability parameter.

Simulation study

To assess the robustness of our method, a variety of design scenarios were considered, including different sample sizes of the non-inferiority trial, amount of available historical information and inter-trial variability.

The comparison between the three methods is summarised in terms of sample size to achieve a given statistical power (Fig. 3), as well as Type I error and power vs. sample size (Fig. 4). The simulated scenario depicted in Fig. 3 refers to the case of 2 historical trials, with sample size of historical trials equal to 50 subjects, and a deviation of $\pm$ 30% from a hypothetical placebo success probability of 0.5 (equivalent to $\omega^2 = 0.1$). In Fig. 4, only the scenario with 50 patients in each historical trial is shown, with supporting data from 2 or 12 historical trials. Similar results were obtained in the remaining scenarios. Moreover, evaluation of different values of inter-trial variability did not yield appreciable differences in the results.

The Bayesian mixed-effects method yielded a definitive and consistent advantage over the standard and non-hierarchical Bayesian techniques, achieving a considerable reduction in the sample size for the new trial (Fig. 3). With respect to the non-hierarchical Bayesian method, for example, about one hundred patients would be saved if 80% statistical power were required in the planned non-inferiority trial. Moreover, it should be noted that sample size would exceed 500 subjects to meet the same criteria based on the standard methods currently used.

As can be noted in Fig. 4, all three methods achieved Type I errors between 1% and 6%. However, for a given sample size, our model yielded a considerable gain in power to detect non-inferiority, which is in accordance with the results of Fig. 3.

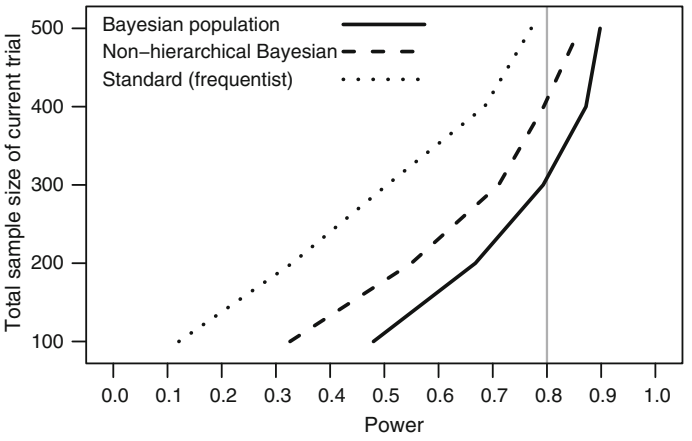

Fig. 3 Relationship between the required sample size for a new trial and statistical power. Bayesian population, non-hierarchical Bayesian and standard methods are compared. Simulations refer to the case of two historical trials using 50 subjects as sample size for historical trials and ±30% deviation from a hypothetical placebo success probability of 0.5

Observe that the integration of historical data using a hierarchical model structure allows increased statistical power not only because estimates of effect sizes are made more precise (due to the larger amount of data contributing to their calculation), but also because heterogeneity of trials is accounted for, which results into a more accurate estimation of the true (i.e., typical) effect sizes.

Discussion

The increasing expectation of patients and health care policy makers on the effectiveness and improved benefit-risk ratio of novel therapeutic agents has prompted the demand for comparative trials in which the superiority of a treatment may not be the primary aim of a protocol. Indeed, there may be circumstances in which a non-inferiority analysis, as opposed to superiority, is unavoidable: fewer side effects, less invasiveness, reduced costs and treatment convenience [37]. This scenario has become more and more relevant with the search for product differentiation, a requirement which needs to be explored already in the early stages of development.

Despite the arguments for the use of active-controlled non-inferiority trials to compare a new drug with existing therapeutic solutions, common concerns are the design requirements and the impossibility to fully characterise the exposure–response relationship. In addition, as indicated previously, the improper or incomplete use of available information concerning placebo response and efficacy of comparators raises not only ethical but also statistical questions regarding the estimates of effect size [3, 38]. We have addressed these issues by developing a general Bayesian method to compare a candidate drug to an established comparator

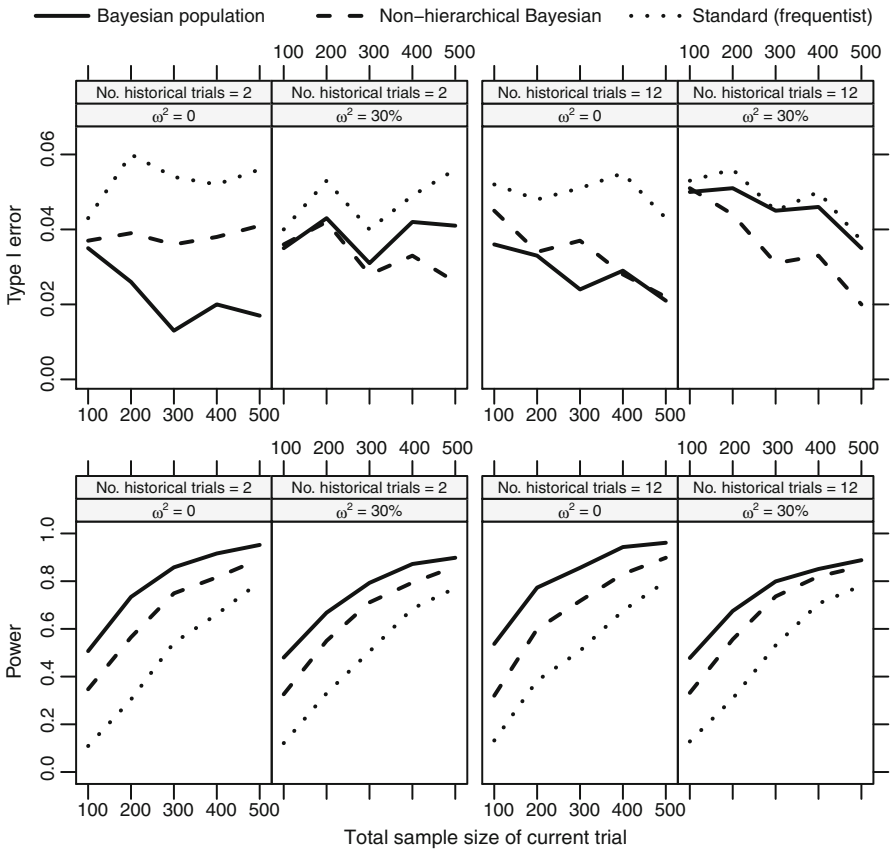

Fig. 4 Type I error (*top*) and power (*bottom*) for Bayesian population, non-hierarchical Bayesian and standard methods. Two (*left panels*) and twelve (*right panels*) historical trials with 50 subjects each are considered, both with and without inter-trial variability (ω^2). Each point is obtained from 1,000 simulations

as well as placebo, in a way which structurally exploits historical information and potentially provides the basis for subsequent evaluation of the exposure–response curve in the target population.

Although it would be possible to elicit estimates of effect sizes from past trials, as it is often done with historical placebo, we do not advocate such a strategy, which may be unpractical and not fully acceptable in a regulatory setting. Instead, we directly incorporate all available information in our model, by means of a hierarchical model structure. Such information may come from past comparator vs. placebo studies as well as from placebo arms in trials with different controls. Our purpose is to "borrow strength" from previous trials while still accounting for their inherent diversity. Moreover, it should be noted that this approach enables the incorporation of pharmacokinetic-pharmacodynamic modelling concepts into the evaluation of non-inferiority trials. An exposure–response curve may be inferred even if a placebo arm for the current non-inferiority trial is not available. On the

other hand, there are also circumstances, such as in oncology and cardiovascular trials, in which patients must be assigned treatments with proven efficacy [1, 39, 40]. Our approach allows estimation of the placebo effect in the non-inferiority trial even if a placebo arm is not present. In this respect, we need not resort to the so-called *constancy assumption*, that is, that the difference between control and placebo effect of a generic past trial is retained in the current clinical setting. Unfortunately, most traditional putative placebo analyses still seem to lean on this widely-criticised assumption [8, 11, 41].

Another common concern relates to the underlying statistical requirements, which cause non-inferiority trials to be "large trials" [42–44] and their tendency, in life-threatening diseases, to expose a large fraction of patients to serious or fatal conditions given the uncertainty of the benefit of the new treatment [45]. However, convincing evidence can be obtained by collecting information from possibly smaller trials [45]. As illustrated here, the integration of all available information from other placebo- or active-controlled trials proves useful in reducing the planned sample size. Therefore, this Bayesian approach circumvents the aforementioned concerns, enabling efficient use of patients participating in non-inferiority trials. Remarkably, our method is also easily applicable to superiority analyses, although these features have not been explored here. For instance, a simple superiority test could be obtained by setting $\mu_1 = 1$ and checking if the probability that T_1 is greater than zero overcomes the threshold P_{cutoff}. The effect retention statistics T_2 would not be necessary, since superiority implies effect retention. Furthermore, the concepts may be easily extended to account for continuous endpoints: e.g., the mean response in a treatment arm may be modelled as a linear or nonlinear function of fixed effects, random trial effects, and measurement error.

The final assessment of the expected effect size relies on two test statistics, namely the probability that the experimental drug is non-inferior to the comparator and that the experimental drug retains a certain portion of the comparator's effect with respect to placebo. Usually, such requirements are stated in terms of the null hypothesis to be rejected. In many cases, indeed, the latter test is implicitly performed as part of the former, that is, the non-inferiority margin is actually estimated by imposing a constraint on the effect retention. This *two-at-the-price-of-one* approach may be misleading and rather cumbersome. The clear-cut approach, adopted in this modelling exercise, is to compute the posterior probability that both non-inferiority and effect retention are guaranteed.

From a regulatory standpoint, it should be noted that despite the acceptance by regulatory authorities, the use of traditional methods may result in biased and/or imprecise estimates of the treatment effect size, which has direct implications for the overall statistical power. The additional drawback currently accepted methods is an inefficient exploitation of sample size properties: regulatory requirements mandate that a non-inferiority trial be repeated, in order to confirm the non-inferiority assessment. An integrated analysis of all available data, including information from past studies, enables increased power to detect small effect sizes as well as to efficiently estimate the sample size for a future non-inferiority trial. The aforementioned limitations are overcome by the approach proposed here and as such can be considered suitable for adoption by regulatory boards.

From the above, it is evident that current standards the analysis of non-inferiority trials represent not only a scientific burden in clinical research. They also prevent further attempts to differentiate new medicines prior to approval. Thus far, this subject has remained beyond the realm of clinical pharmacology, with compounds progressing without clarity about the underlying exposure–response curve and consequently about the expected performance of the novel treatment. A paradigm shift is required to address unmet medical needs. Non-inferiority comparisons are useful and clinically meaningful. However, the level of evidence and methods required for such an evaluation cannot continue to neglect the need to accurately use existing information, optimise study design and characterise the exposure–response relationships of these compounds. Clinical differentiation and quantitative assessment of the characteristics of a medicinal product ought to result from an integrated analysis of available evidence and as such, efforts must consider historical information.

Conflict of interest None.

References

1. Kaul S, Diamond GA (2006) Good enough: a primer on the analysis and interpretation of noninferiority trials. Ann Intern Med 145:62–69
2. Carroll KJ (2006) Active-controlled, non-inferiority trials in oncology: arbitrary limits, infeasible sample sizes and uninformative data analysis. Is there another way? Pharm Stat 5:283–293
3. Garattini S, Bertelé V (2007) Non-inferiority trials are unethical because they disregard patients' interests. Lancet 370:1875–1877
4. Simon R (2000) Are placebo-controlled clinical trials ethical or needed when alternative treatment exists? Ann Intern Med 133:474–475
5. Food and Drug Administration (2010). Guidance for Industry, Non-Inferiority Clinical Trials. March 2010. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf
6. European Medicines Agency. Guideline on the choice of the non-inferiority margin. Committee for Proprietary Medicinal Products. Doc. Ref. EMEA/CPMP/EWP/2158/99. July 2005. http://www.ema.europa.eu/pdfs/human/ewp/215899en.pdf
7. Christensen E (2007) Methodology of superiority vs. equivalence trials and non-inferiority trials. J Hepatol 46:947–954
8. Hung HMJ, Wang SJ, Tsong Y, Lawrence J, O'Neill R (2003) Some fundamental issues with non-inferiority testing in active controlled trials. Stat Med 22:213–225
9. Durrleman S, Chaikin P (2003) The use of putative placebo in active control trials: two applications in a regulatory setting. Stat Med 22:941–952
10. Hasselblad V, Kong DF (2001) Statistical methods for comparison to placebo in active-control trials. Drug Inf J 35:435–449
11. D'Agostino RB, Massaro JM, Sullivan LM (2003) Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. Stat Med 22:169–186
12. Goodman S (1999) Toward evidence-based medical statistics. Part 1: the *P* value fallacy. Ann Intern Med 130:995–1004
13. George A, Rubin G (2003) A systematic review and meta-analysis of treatments for impetigo. Br J Gen Pract 53:480–487

14. Koning S, Verhagen AP, van Suijlekom-Smit LWA, Morris AD, Butler C, van der Wouden JC (2009) Interventions for impetigo. The Cochrane Collaboration. http://www.thecochranelibrary.com
15. Simon R (1999) Bayesian design and analysis of active control clinical trials. Biometrics 55:484–487
16. Christensen OB, Anehus S (1994) Hydrogen peroxide cream: an alternative to topical antibiotics in the treatment of impetigo contagiosa. Acta Derm-Venerologica 74:460–462
17. Eells LD, Mertz PM, Piovanetti Y, Pekoe GM, Eaglstein WH (1986) Topical antibiotic treatment of impetigo with mupirocin. Arch Dermatol 122:1273–1276
18. Gilbert M (1989) Topical 2% mupirocin versus 2% fusidic acid ointment in the treatment of primary and secondary skin infections. J Am Acad Dermatol 20:1083–1087
19. Gould JC, Smith JH, Moncur H (1984) Mupirocin in general practice: a placebo-controlled trial. Royal society of medicine: international congress and symposium series 180:85–93
20. Koning S, van Suijlekom-Smit LWA, Nouwen JL, Verduin CM, Bernsen RMD, Oranje AP, Thomas S, van der Wouden JC (2002) Fusidic acid cream in the treatment of impetigo in general practice: a double blind randomised placebo controlled trial. BMJ 324:203–206
21. Moraes Barbosa AD (1986) Comparative study between topical 2% sodium fusidate and oral association of chloramphenicol/neomycin/bacitracin in the treatment of staphylococcic impetigo in newborn. Arq Bras Med 60:509–511
22. Morley PAR, Munot LD (1988) A comparison of sodium fusidate ointment and mupirocin ointment in superficial skin sepsis. Curr Med Res Opin 11:142–148
23. Park SW, Wang HY, Sung HS (1993) A study for the isolation of the causative organism, antimicrobial susceptibility tests and therapeutic aspects in patients with impetigo. Korean J Dermatol 31:312–319
24. Rojas R, Eells L, Eaglstein W, Provanetti Y, Mertz PM, Mehlisch DR, et al (1985) The efficacy of Bactroban ointment and its vehicle in the treatment of impetigo: a double-blind comparative study. Proceedings of an international symposium, Nassau, Bahama Islands, 21–22 May 1984, pp 96–102
25. Ruby RJ, Nelson JD (1973) The influence of hexachlorophene scrubs on the response to placebo or penicillin therapy in impetigo. Pediatrics 52:854–859
26. Sutton JB (1992) Efficacy and acceptability of fusidic acid cream and mupirocin ointment in facial impetigo. Curr Ther Res 51:673–678
27. Vainer G, Torbensen E (1986) Treatment of impetigo in general practice: comparison of 3 locally administered antibiotics. Ugeskrift laeger 148:1202–1206
28. White DG, Collins PO, Rowsell RB (1989) Topical antibiotics in the treatment of superficial skin infections in general practice – a comparison of mupirocin with sodium fusidate. J Infec 18:221–229
29. Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Anal 1:515–533
30. Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. Wiley, Chichester
31. International Conference on Harmonisation, Topic E 10 (Choice of control group in clinical trials). Note for guidance on choice of control group in clinical trials. Doc. Ref. CPMP/ICH/364/96. January 2001. http://www.ema.europa.eu/pdfs/human/ich/036496en.pdf
32. Hung HMJ, Wang SJ, O'Neill R (2005) A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. Biom J 47:28–36
33. Lange S, Freitag G (2005) Choice of delta: requirements and reality—results of a systematic review. Biom J 47:12–27
34. Lunn DJ, Thomas A, Best N, Spiegelhalter DJ (2000) WinBUGS—a Bayesian modelling framework: concepts, structure and extensibility. Stat Comput 10:325–337
35. R Development Core Team (2008) *R:* A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
36. Sutton A, Abrams KR, Jones D, Sheldon T, Song F (2000) Methods for meta-analysis in medical research. Wiley, Chichester
37. Pocock SJ (2003) The pros and cons of noninferiority trials. Fund Clin Pharmacol 17:483–490
38. Garattini S (2005) Designing the most favourable study design. Eur J Clin Pharmacol 61:85–86
39. World Medical Association (2008) 59th General Assembly, Declaration of Helsinki: ethical principles for medical research involving human subjects. Seoul, October 2008. http://www.wma.net/en/30publications/10policies/b3/17c.pdf
40. Pater C (2004) Equivalence and noninferiority trials—are they viable alternatives for registration of new drugs? Curr Control Trials Cardiovasc Med 5:8

41. Wang SJ, Hung HMJ, Tsong Y (2002) Utility and pitfalls of some statistical methods in active controlled clinical trials. Control Clin Trials 23:15–28
42. Peto R, Collins R, Gray R (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 48:23–40
43. Ellenberg SS, Foulkes MA (1994) The utility of large, simple trials in the evaluation of AIDS treatment strategies. Stat Med 13:408–415
44. Yusuf S, Collins R, Peto R (1984) Why do we need some large, simple randomized trials? Stat Med 3:971–980
45. Palmer CR (2002) Ethics, data-dependent designs, and the strategy of clinical trials: time to start learning-as-we-go? Stat Methods Med Res 11:381–402