Article

CLINICAL TRIALS

Clinical Trials
2017, Vol. 14(3) 277–285

DOI: 10.1177/1740774517699409
journals.sagepub.com/home/ctj
SAGE

Using phase II data for the analysis of phase III studies: An application in rare diseases

Simon Wandel[1], Beat Neuenschwander[1], Christian Röver[2] and Tim Friede[2]

Abstract

Background: Clinical research and drug development in orphan diseases are challenging, since large-scale randomized studies are difficult to conduct. Formally synthesizing the evidence is therefore of great value, yet this is rarely done in the drug-approval process. Phase III designs that make better use of phase II data can facilitate drug development in orphan diseases.

Methods: A Bayesian meta-analytic approach is used to inform the phase III study with phase II data. It is particularly attractive, since uncertainty of between-trial heterogeneity can be dealt with probabilistically, which is critical if the number of studies is small. Furthermore, it allows quantifying and discounting the phase II data through the predictive distribution relevant for phase III. A phase III design is proposed which uses the phase II data and considers approval based on a phase III interim analysis. The design is illustrated with a non-inferiority case study from a Food and Drug Administration approval in herpetic keratitis (an orphan disease). Design operating characteristics are compared to those of a traditional design, which ignores the phase II data.

Results: An analysis of the phase II data reveals good but insufficient evidence for non-inferiority, highlighting the need for a phase III study. For the phase III study supported by phase II data, the interim analysis is based on half of the patients. For this design, the meta-analytic interim results are conclusive and would justify approval. In contrast, based on the phase III data only, interim results are inconclusive and require further evidence.

Conclusion: To accelerate drug development for orphan diseases, innovative study designs and appropriate methodology are needed. Taking advantage of randomized phase II data when analyzing phase III studies looks promising because the evidence from phase II supports informed decision-making. The implementation of the Bayesian design is straightforward with public software such as R.

Keywords

Drug development in rare diseases, phase III studies, Bayesian statistics, meta-analysis

Introduction

Clinical research in orphan diseases is challenging. It is often impossible or unethical to conduct large-scale randomized controlled trials, which implies that only limited evidence is available for decision-making. Also, shortcomings in the methodological approaches to evaluate medical products in rare diseases have been identified.[1] While these problems have been recognized for some time,[2] only in the past few years strong efforts have been made to address them. Examples include the draft guidance by the Food and Drug Administration (FDA) for drug development in rare diseases[3] and the latest funding scheme for rare diseases by the European Union's Horizon 2020 research program.[4] These activities have led to intensified rare diseases research and drug development by pharmaceutical companies.[5]

With regard to the drug-approval process, some flexibility on study designs and endpoints has been observed for drugs with an orphan indication.[6,7] Surprisingly, however, a formal combination of the evidence (e.g. a meta-analysis) is rarely presented in

[1]Novartis Pharma AG, Basel, Switzerland
[2]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

Corresponding author:
Simon Wandel, Novartis Pharma AG, Postfach, 4002 Basel, Switzerland.
Email: simon.wandel@novartis.com

approval dossiers. Typically, efficacy is assessed based on confirmatory trials only, meaning that other evidence (such as phase II studies) is viewed as supportive only. This poses a problem to both, regulators in charge of approving drugs and companies developing them, since it limits the evidence base for a quantitative assessment of the treatment effect. Furthermore, the combination of data reveals its power particularly in rare diseases: on one hand, because data are limited, and on the other hand, because the non-confirmatory trials typically compose a large fraction of all patients in the development program. The latter may be different for non-rare diseases, where the majority of patients is usually enrolled in the large-scale phase III trials.

These challenges call for approaches to study design and analysis that allow a more efficient use of the available data, as stipulated, for example, in the 21st Century Cures Act.[8] The nature of the problem lends itself to the Bayesian approach. Its usefulness when meta-analyzing few (small) studies has been discussed elsewhere.[9,10] Here, we extend the idea to incorporate existing evidence for the parameter of interest, the treatment effect corresponding to the phase III study, via a meta-analysis. This is based on concepts discussed by Spiegelhalter et al.,[11] Neuenschwander et al,[12,13] Schmidli et al.,[14] and some ideas in Gerß and Köpcke.[15]

The article is organized as follows. We first describe the statistical methodology, then illustrate the design using data from an FDA approved drug, and conclude with a discussion.

Methods

Hierarchical models

Hierarchical models (HM) are widely used when data are available from more than one trial. The models have two components: a data model and a parameter model. Data Y_j from trial $j = 1, \ldots, J$ follow a distribution F parameterized by trial-specific parameters θ_j

$$Y_j|\theta_j \sim F(\theta_j) \tag{1}$$

and trial parameters θ_j follow a distribution G

$$\theta_j|\eta \sim G(\eta) \tag{2}$$

Inference for trial parameters can be done in a classical or Bayesian way. The simplest HM assumes (approximately) normal data. Often, the Y_j are parameter estimates rather than individual data. For this case, the normal–normal hierarchical model (NNHM) is widely used

$$Y_j|\theta_j \sim N(\theta_j, s_j^2) \tag{3}$$

and

$$\theta_j|\mu, \tau \sim N(\mu, \tau^2) \tag{4}$$

For fixed standard errors s_j and known (assumed) τ, classical and Bayesian conclusions for μ and trial parameters θ_j are the same if a non-informative (improper) prior for μ is used. For precision (inverse-variance) weights w_j, total precision w_+, and shrinkage parameters B_j

$$w_j = \frac{1}{s_j^2 + \tau^2}, \qquad w_+ = \sum_{j=1}^{J} w_j, \quad B_j = \frac{s_j^2}{s_j^2 + \tau^2} \tag{5}$$

the posterior distribution of μ based on $Y = (Y_1, \ldots, Y_J)$ is

$$\mu|Y, \tau \sim N(\sum w_j Y_j / w_+ , 1/w_+) \tag{6}$$

The posterior distributions of the trial parameters θ_j are

$$\theta_j|Y, \tau \sim N(B_j\hat{\mu} + (1 - B_j)Y_j, B_j(\tau^2 + B_j/w_+)) \tag{7}$$

where $\hat{\mu}$ is the posterior mean in equation (6). Classical analogues to the posterior means and standard deviations are maximum-likelihood estimates and their standard errors. The special cases of complete pooling and stratification arise for $\tau = 0$ and $\tau = \infty$, respectively.

Intermediate values of τ lead to different degrees of information sharing across trials, with the desirable properties one expects from an approach aiming to improve inference by borrowing information from similar trials:

- The HM shrinks the trial estimates toward the estimate of μ, which acts as a safeguard against over-interpreting extreme (good or bad) trial results. Shrinkage depends on trial size and between-trial heterogeneity. For large trials (small s_j), shrinkage is small, and notable shrinkage is only possible if τ is of small to moderate size.
- The HM improves precision. Since

$$B_j\left(\tau^2 + \frac{B_j}{w_+}\right) = s_j^2 - s_j^2 B_j\left(1 - \frac{w_j}{w_+}\right) \tag{8}$$

the variance in equation (7) is always smaller than the variance s_j^2 of Y_j.

Between-trial heterogeneity

The degree of between-trial heterogeneity (standard deviation τ in equation (4)) for the parameters $\theta_1, \ldots, \theta_J$ depends on the parameter scale and the outcome standard deviation σ for one observation unit (e.g. one subject or one event). Table 1 shows four typical heterogeneities and respective τ values for $\sigma = 2$, which is often used as the reference standard deviation

Table 1. Classification of between-trial heterogeneity with 97.5% quantile to median ratio for risk ratio (RR) trial parameters; σ is the outcome standard deviation, and τ is the between-trial standard deviation on the log(RR) scale.

Heterogeneity (σ/τ)	τ (if $\sigma = 2$)	$RR_{97.5\%}/RR_{50\%}$
Large (2)	1	7.10
Substantial (4)	0.5	2.66
Moderate (8)	0.25	1.63
Small (16)	0.125	1.28

for normal approximations of binomial, count, and survival data.[11] For the four heterogeneities, Table 1 shows the range of parameter values expressed as the ratio between the 97.5% quantile and the median. For example $\tau = 1$ implies a ratio of 7.1, which is clearly large and will be rare in practice. These values will also be helpful in the application, where we analyze the data using log-risk-ratios, $\log(\mathrm{RR_j}) = \log(p_{j,T}/p_{j,C})$ and $p_{j,T}$, $p_{j,C}$ are the response rate in the treatment and control group, respectively.

For the common case of few trials, the size of between-trial heterogeneity is usually highly uncertain because τ cannot be inferred well from the data. Therefore, it is important to use prior distributions covering plausible τ values. Half-normal, half-Cauchy, and half-t distributions have been suggested in this context.[11,16,17] For the log-risk ratios used in the application, we will consider half-normal distributions[11] with scale parameters 0.5 and 1, which have medians (95%-intervals) equal to 0.34 (0.016, 1.12) and 0.67 (0.031, 2.24), respectively. Since $\tau = 1$ represents large heterogeneity, both priors are weakly informative, covering small to large heterogeneity and leaving small probabilities to unrealistically large heterogeneities, whereby the latter prior (with median 0.67) is rather conservative. For these priors, the 97.5% quantile to median ratio for risk ratio (RR) trial parameters is 2.98 and 8.89, respectively (see Appendix 1).

Meta-analytic-predictive prior

When designing a new trial with parameter $\theta_\star$, the predictive distribution based on previous data $Y_1, \ldots, Y_J$ constitutes the prior distribution for the new trial. This is known as the meta-analytic-predictive (MAP) prior[11,12,14]

$$\theta_\star | Y_1, \ldots, Y_J \tag{9}$$

For the NNHM with known τ

$$\theta_\star | Y_1, \ldots, Y_J, \tau \sim N(\hat{\mu}, \tau^2 + 1/w_+) \tag{10}$$

which follows from equation (7) by adding the new trial (with no data) to the model, that is, $s_\star = \infty$ and $B_\star = 1$.

Analysis for new trial

Eventually, after the new data $Y_\star$ have been observed, inference for $\theta_\star$ can be done in two ways:

- MAP The MAP approach formally combines the prior (equation (9)) with $Y_\star$ in a standard Bayesian way.
- MAC The meta-analytic-combined (MAC) approach does not require a prior distribution for $\theta_\star$. It simply infers $\theta_\star$ at the end of the new trial by a meta-analysis of historical and new data, resulting in

$$\theta_\star | Y_1, \ldots, Y_J, Y_\star \tag{11}$$

Importantly, MAP and MAC give identical results.[14] The MAP approach is technically more involved because MAP priors (equation (9)) do not follow standard distributions and are typically heavy-tailed. This complicates the Bayesian analysis with $Y_\star$ at the end of the trial, which can be addressed via mixture approximations.[14] However, even if a MAC analysis will usually be the method of choice and easy to perform with meta-analytic software, the MAP prior plays an important role: it quantifies prior information at the design stage, which may be required in the trial protocol.

Effective sample sizes

In many applications, the use of appropriately discounted prior information, which accounts for between-trial heterogeneity, will lead to smaller trials, unless heterogeneity is large. The prior information can be expressed as an equivalent approximate prior effective sample size (ESS). In our setting, we are interested in $ESS_\star$, the prior ESS of the MAP prior (equation (9)). Various approaches to ESS have been proposed;[12,18–21] they are similar in the sense that they relate the ESS to the precision (inverse of variance) of the prior distribution.

Here, we will use an approximate two-variances approach which requires the following: the variance $V_\star$ of the analysis of interest, for which the $ESS_\star$ is unknown, and the variance V_0 of a simpler analysis (e.g. a meta-analysis with $\tau = 0$) with known ESS_0. Assuming that $ESSs$ are approximately proportional to precisions, the ESS of interest is

$$ESS_\star = ESS_0 \times \frac{V_0}{V_\star} \tag{12}$$

In our case, $V_\star$ will be the variance of the MAP prior (equation (9)), whereas V_0 will be the one from the analysis assuming no between-trial heterogeneity ($\tau = 0$).

Case study

We now illustrate how to use phase II data for the design and analysis of a phase III study. The design

Table 2: Data of phase II and phase III studies.

	Study (phase)			
	4 (II)	5 (II)	6 (II)	7 (III)
Objective	Efficacy and safety	Efficacy and safety	Efficacy and safety	Efficacy and safety
Design	Three-arm randomized	Two-arm randomized	Three-arm randomized	Two-arm randomized
Location	Africa	Europe	Pakistan	Europe and Africa
Product	G: 0.15%, 0.05%; A: 3%	G: 0.15%; A: 3%	G: 0.15%, 0.05%; A: 3%	G: 0.15%; A: 3%
Regimen	1	1	2	1
Study period (months)	4/90–5/92 (25)	12/90–5/92 (18)	5/91–10/92 (18)	9/92–9/94 (25)
Total cure rate, day 14 (%)				
Zirgan	19/23 (82.6)	15/18 (83.3)	31/36 (86.1)	74/84 (88.1)
Acyclovir	16/22 (72.7)	12/17 (70.6)	27/38 (71.1)	73/80 (91.3)

Regimen: 1 = 1 drop 5x/day until ulcer healed, then 1 drop 3x/day for 7 days; 2 = 1 drop 5x/day for 10 days.

relies on the methodology of the previous section and additional considerations such as practical feasibility and regulatory requirements. Data from three phase II and one phase III trial on Zirgan (0.15% gel) for the treatment of acute herpetic keratitis will be used in the case study. All analyses were conducted in R[22] with the package bayesmeta[23] (see Appendix 1 for code).

Background

Herpetic keratitis is an inflammatory condition of the eye caused by an outbreak of the herpes simplex virus (HSV).[24,25] It can have serious consequences and remains the leading cause of corneal blindness in the industrialized world.[26,27] With as few as 1.5 million people affected world wide,[28] it has been classified as an orphan indication by the FDA [29] and the European Medicines Agency.[30]

In 2009, the FDA approved Zirgan for the treatment of herpetic keratitis (dendritic ulcers).[31] To discuss all details of the approval is beyond the scope of this application (see the publicly available documents[32]). However, a few points are noteworthy. Most importantly, from the files,[29,32] it appears that approval was based on a retrospective analysis of the four relevant studies, three phase II and one phase III study. Retrospective means that the sponsor submitted the results of the studies after they were conducted, rather than seeking the agency's advice beforehand. Subsequently, this led to discrepancies between the sponsor's and FDA's primary analyses, including changes of the population, of the endpoint, and from superiority to non-inferiority.

The reasons behind this rather unusual approach to approval are not entirely clear. One explanation may be that the original manufacturer (Théa of France) did not intend to bring Zirgan to the US market on its own; rather, it sold the license for the US market to Sirion Therapeutics in 2007 which then initiated the submission. This and the fact that the clinical studies were already conducted in the 1990s may explain why no early discussions with the FDA took place.

Our goal here is not to reconstruct the approval history in detail. Rather, we will use the example to discuss an alternative, more efficient statistical approach toward approval, based on the following design specifications in the non-inferiority setting: cure rate at day 14 as endpoint, dendritic and geographic ulcers as population, and an absolute non-inferiority margin of 12% points. Furthermore, we will use the RR to quantify the treatment effect.

In the following, we present the evidence available at the hypothetical end-of-phase II meeting, a potential phase III trial and approval strategy, and the results of the actual phase III trial.

Hypothetical end-of-phase II meeting

Three randomized phase II studies[33] were conducted between April 1990 and October 1992 (Table 2). The studies were similar, with the only minor difference being the treatment regimen in study 6. For simplicity, we assume that this difference is not relevant for the clinical outcome.

We now turn the clock back and assume we are in the situation of an end-of-phase II meeting. We assume that the sponsor would agree to a non-inferiority analysis of Zirgan versus Acyclovir (the standard of care) with the primary endpoint being cure rate at day 14. Actually, setting the non-inferiority margin proved to be difficult. For cure rate at day 14, the FDA determined two effect sizes[34,32] M1: 14% and 18%. The latter implies an absolute non-inferiority margin of 12% points when retaining one-third of the effect. We assume here that this margin had been agreed to.

At this stage, it is interesting to perform a non-inferiority analysis (Zirgan versus Acyclovir) of the phase II data. If the evidence were overwhelming, it would be fair to ask whether a phase III study were required, or if approval could be granted based on the phase II data only.

Our interest is the phase III treatment effect. However, since no phase III data are available yet, the phase III treatment effect corresponds to the predicted treatment effect $\theta_\star$ from the phase II studies. The underlying statistical model is the NNHM (equations (3) and (4)), with study-specific estimates of the log-risk ratios $Y_j = \log(\mathrm{RR}_j)$ and standard errors

$$s_j = \sqrt{1/r_C - 1/n_C + 1/r_T - 1/n_T} \qquad (13)$$

where n and r denote the number of patients and responders. This requires a transformation to the risk difference scale and a sensible prior distribution for the between-trial heterogeneity parameter τ (the prior for μ will be non-informative).

The first point is straightforward. For a response rate p_C in the control group and a pre-defined non-inferiority margin $m = p_C - p_T$, the transformation is given by the definition of the RR; for $p_C = 0.9$ (the cure rate observed in active-controlled Acyclovir studies that did not use Zirgan[32]) and margin $m = 0.12$, non-inferiority holds if $\mathrm{RR}_{T:C} \geq 0.867$. To declare non-inferiority, we require 97.5% posterior probability that $\theta_\star \geq \log(0.867)$, which is equivalent to the lower bound of the central 95% credible interval being equal to (or above) that threshold. This choice is motivated by the analysis of FDA's statistical reviewer, who used a 95% confidence interval to assess non-inferiority.

For the τ prior, we use $\tau \sim HN(0.5)$; here, $HN(scale)$ is the half-normal distribution: for $x \sim N(0, scale^2)$, $y = |x|$ follows a $HN(scale)$ distribution. The $HN(0.5)$ prior is centered at moderate to substantial heterogeneity (median = 0.34) and covers small to large heterogeneity (95% interval = (0.016;1.12)); see Table 1. Notably, one may perform a sensitivity analysis using a prior which favors larger between-trial heterogeneity, for example, $\tau \sim HN(1)$, or empirical priors as proposed in Turner et al.[35]

The meta-analysis of the phase II data is shown in Figure 1, where the data, study-specific (stratified) RR_j, the population mean μ and the predicted effect $\theta_\star$ are shown. The reference line is drawn at the non-inferiority margin ($\mathrm{RR}_{T:C} = 0.867$ for $p_C = 0.9$). The posterior for τ indicates small between-trial heterogeneity, with median 0.12 (95% interval 0.00 to 0.51).

The meta-analysis provides evidence for non-inferiority. If μ were the parameter of interest, an almost conclusive statement would follow: the lower bound of the 95% interval is just below the non-inferiority margin. In fact, $p(\mu \geq \log(0.867)) = 97.1\%$, very close to 97.5%. However, the parameter $\theta_\star$ in the phase III trial is of interest. For this parameter, the evidence for non-inferiority is weaker, but still substantial: $p(\theta_\star \geq \log(0.867)) = 92.0\%$. These results are based on the normal approximation for the log-risk ratio, for which the accuracy may be questioned. However, when using a model with binomial likelihoods instead, results are essentially identical: $p(\mu \geq \log(0.867)) = 96.9\%$ and $p(\theta_\star \geq \log(0.867)) = 91.7\%$. This shows that the normal approximation works well in this example.

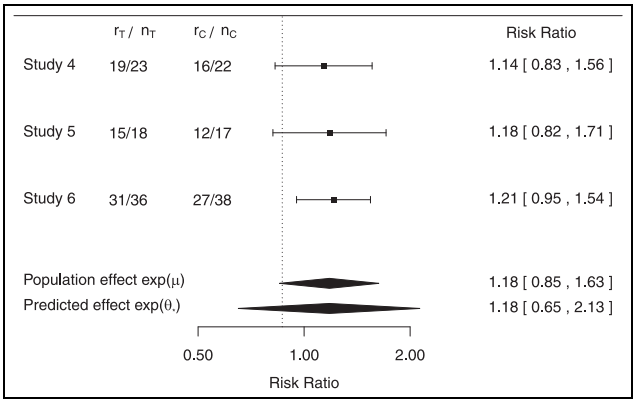

Figure 1. Data and results at end-of-phase II meeting.

Phase III study and proposed strategy for approval

Designing a phase III study that allows to assess non-inferiority in combination with the available evidence is desirable. This will not only allow to run a smaller study but also provide a treatment effect estimate based on all relevant evidence. However, regulators may have good reasons to argue that a smaller study may provide insufficient information for approval, especially to assess the safety and risk/benefit ratio.

We now discuss the design of a phase III study (study 7) which uses phase II data and allows for seeking approval based on an interim analysis. Depending on negotiations with regulators, a post-approval commitment to run the study to its end (even if approval is granted at interim) may be required. However, such negotiations will always be case-specific, highlighting the importance of early discussions with regulators. Nevertheless, the option to seek approval based on a positive interim analysis seems attractive for this case study. Since the endpoint is evaluated at day 14, there will be a small time window between the last patient enrolled for the interim analysis and the actual data read-out and analysis. With an anticipated recruitment period of 2 years, such a strategy could result in a markedly earlier approval.

When seeking approval based on interim results, the information fraction for the interim analysis becomes a key design aspect. We will assume that the interim analysis is conducted after 50% of the patients have been evaluated. For the sample size, in order to align with the actual study as originally conducted, we will assume $n_C = n_T = 80$. This results in interim sample sizes $\widetilde{n}_C = \widetilde{n}_T = 40$.

It is also important to understand how much phase II information is borrowed (which depends on the between-trial heterogeneity) when inferring the phase III effect. Using the variance ratio approach, the ESS is 14.

Operating characteristics

We evaluate the operating characteristics (type-I error rate and power) of the design and compare them to a phase III design ignoring the phase II data. For different fixed response rates p_C (reflecting potential differences between phase II and phase III) and fixed treatment differences δ, we used R[22] to simulate interim and final data

$$r_C^{k,int} \sim Bin(n_C^{int}, p_C)$$
$$r_T^{k,int} \sim Bin(n_T^{int}, p_C + \delta)$$
$$r_C^{k,final} \sim r_C^{k,int} + Bin(n_C^{final} - n_C^{int}, p_C)$$
$$r_T^{k,final} \sim r_T^{k,int} + Bin(n_T^{final} - n_T^{int}, p_C + \delta)$$

for $k = 1, \ldots, 10,000$. The data were analyzed with the package bayesmeta,[23] either incorporating the phase II data (meta-analysis) or not (phase III study alone). The operating characteristics are presented in Table 3, which shows two probabilities: the probability to be successful at the final analysis (regardless of the outcome of the interim analysis) and the probability to be successful both at the interim and the final analysis.

The gain in power for the proposed design can be substantial. For example, for $p_C = 0.7$ (the Acyclorivr cure rate observed in phase II) and $\delta = 0.06$, the power is 87% versus 66%. The power gain is even larger at interim (70% versus 35%). When $p_C = 0.9$ (the cure rate observed in active-controlled Acyclovir studies that did not use Zirgan[32]) and $\delta = 0$, the power is 87% versus 79%, and 68% versus 48% at interim. The larger increase in power at interim is remarkable and due to the highly consistent phase II results, which suggested superiority of Zirgan.

The gain in power, however, comes at the price of an increased type-I error rate. Strict type-I error rate control cannot be guaranteed.[36] For example, for $p_C = 0.7$ and $p_C = 0.9$, the type-I error rates are 6% versus 1% and 8% versus 3%; this increase is not dramatic. In general, a potential inflation of type-I error needs to be discussed with regulators during the design phase. If it is of concern, robust prior distributions could be considered.[13,14] Such heavy-tailed priors ensure that historical information will essentially be ignored if the new and historical data conflict with each other.

Actual phase III data and analysis

The actual data observed in the phase III study are only available for the final analysis. In order to reconstruct an interim analysis using half of the patients, we use an interim sample size of 40 per arm. Furthermore, we choose the number of responders such that observed response rate at interim is close to the observed response rate at the final analysis (see Figure 2).

The results are presented in Figure 2. The interim analysis based on all data (meta-analysis) allows to declare non-inferiority. Note that non-inferiority is claimed based on the parameter corresponding to study 7 ($\theta_\star$) incorporating the evidence from studies 4, 5, and 6. However, the evidence from the phase III study alone is insufficient to declare non-inferiority at interim because the 95% interval includes the non-inferiority margin.

As mentioned before, the idea would be to gain approval with the interim phase III data supported by phase II via the meta-analysis, assuming that other data (such as safety) are also favorable. Yet, depending on negotiations with regulators, the study may still run to its end, allowing a more robust evaluation of the effect at the final analysis. The results for the final analysis are also shown in Figure 2. For the meta-analysis,

Table 3. Operating characteristics for phase II/III (meta-analysis) and phase III alone.

	$\delta = p_T - p_C$				
p_C	–0.12	–0.06	0.0	0.06	0.12
0.70	6 (3)	25 (15)	56 (39)	87 (70)	98 (90)
	1 (0)	8 (3)	30 (12)	66 (35)	93 (67)
0.75	7 (4)	26 (16)	61 (44)	91 (75)	100 (94)
	1 (0)	10 (4)	36 (16)	76 (44)	98 (78)
0.80	7 (4)	29 (18)	68 (49)	94 (80)	100 (97)
	2 (1)	13 (5)	46 (22)	87 (57)	100 (90)
0.85	7 (4)	32 (19)	76 (55)	98 (88)	100 (100)
	3 (1)	17 (7)	60 (31)	95 (72)	100 (99)
0.90	8 (4)	38 (24)	87 (68)	100 (98)	–
	3 (1)	26 (11)	79 (48)	100 (94)	–

Percentages presented: probability for success at final (probability for success at interim and final). The first row corresponds to the meta-analysis, the second row to the analysis of the phase III study alone.

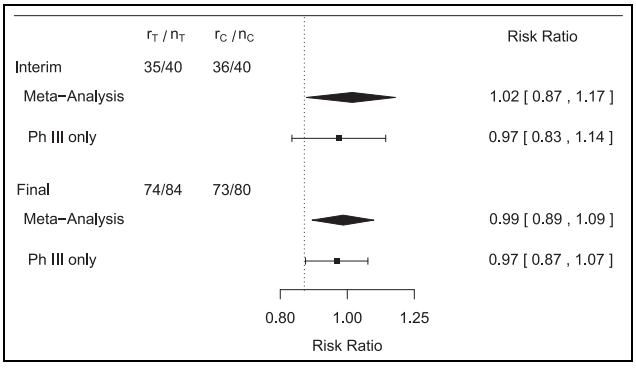

Figure 2. Data and results for interim and final analysis in Phase III.

the interval for the RR becomes narrower and still excludes the non-inferiority margin, thus confirming the interim result. The analysis using the phase III study leads to a lower bound of the interval (0.870) which is just above the non-inferiority threshold 0.867, also allowing to conclude non-inferiority.

Finally, results for τ indicate small between-trial heterogeneity at the interim and the final analysis. The posterior median (95% interval) is 0.12 (0.00–0.41) for the interim and 0.13 (0.00–0.43) for the final analysis. This supports the consistency of the results across all studies.

Discussion

Here, we presented a simple, yet attractive design in rare diseases using phase II data in phase III studies. We illustrated it for binary endpoints using the normal approximation for the log-risk ratio. However, the extension to other models, for example, with binomial likelihoods for both groups, or to other endpoints is straightforward.

The proposed approach uses the phase II data prospectively, which has obvious advantages. First, fewer patients are required in the phase III study, and second, all available evidence is combined. And third, due to the nature of the approach, extreme results will be pulled toward the population mean. The Zirgan case study used to illustrate the design is built on real data as submitted to the FDA. However, the FDA approved Zirgan for a different indication (dendritic ulcers only) and endpoint (cure rate at day 7) than those used in our case study.

Of course, as with any design, all stakeholders need to be convinced. It may be argued that the case study is quite atypical since phase II studies are often not randomized in orphan diseases. This, however, becomes a self-fulfilling prophecy: if evidence from randomized phase II studies is only considered supportive, there is little motivation to perform them. However, if data from randomized phase II studies could be used, this would make them more attractive. It is therefore important that patient groups, regulators and sponsors consider such designs.

Other designs have been proposed before, and an excellent overview is given in Korn et al.[37] Some have been implemented in practice, for example, the historical control monotherapy design proposed by French et al.[38] This design was used successfully, resulting in the approval of Aptiom (eslicarbazepine acetate) for the treatment of partial-onset seizures.[39,40] Other examples include N-of-1 trials,[41] global studies,[42] or basket trials, for example, the B2225 study for Imatinib.[43]

Importantly, the approach that we presented is useful in situations where it is feasible to conduct randomized phase II and phase III studies of reasonable size. For many rare diseases, this may actually be possible. For example, in their review of rare disease terminology and definition, Richter et al.[44] found that the majority of the investigated countries define a rare disease starting at a prevalence of 50/100,000. Similarly, about half of the orphan indications for which drugs were authorized by the European Medicines Agency from 2000 to 2015 had a prevalence between 1/10,000 and 5/10,000.[45] However, for other situations, such as in ultra-rare diseases, alternative approaches may be more appropriate. These include prior elicitation from experts following the SHeffield ELicitation Framework (SHELF),[46] informative priors for the control group based on observational (e.g. registry) data, combining randomized and non-randomized evidence,[47] or N-of-1 trials.[41]

It is also worth mentioning that recent initiatives to improve the drug development process send encouraging signals that a better use of the evidence is welcomed. Important directions are given in the 21st Century Cures Act,[8] which encourages the FDA to further evaluate the use of Bayesian methodology and non-randomized evidence. Furthermore, calls have been made to make the drug approval process more continuous and flexible to account for evidence as it accumulates.[48] The European Medicines Agency has also initiated various working groups.

It is clear that we only considered a small portion of the drug approval process. Efficacy plays a unique role when seeking approval, but other measures are also important. Safety is critical, and additional evidence may be required to assess long-term risks. However, this can often be achieved as a post-approval requirement in the form of non-randomized open-label studies. This approach has the advantage that patients have early access to the treatment while additional data are collected.

The proposed approach has limitations. The potential increase in type-I error needs to be considered and may require design modifications, including robust meta-analytic models.[14] Likewise, for a non-inferiority design, one may consider to directly model the risk

difference and use a meta-analytic approach on this scale.[49] However, most applications will be superiority trials, for which relative measures such as RRs or odds ratios are common.

The motivation of this article was not to challenge FDA's decision. On the contrary, only due to the many publicly available FDA documents, we were able to use this insightful example. We hope that it will facilitate the implementation of the proposed design in practice.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr Wandel and Dr Neuenschwander are employed by Novartis Pharma AG, Basel, Switzerland.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has received funding from the EU's 7th Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013-602144 with project title (acronym) "Innovative Methodology for Small Populations Research" (InSPiRe).

References

1. Unkel S, Röver C, Stallard N, et al. Systematic reviews in paediatric multiple sclerosis and Creutzfeldt-Jakob disease exemplify shortcomings in methods used to evaluate therapies in rare conditions. *Orphanet J Rare Dis* 2016; 11: 16.
2. Orphan Drug Act (ODA). https://www.gpo.gov/fdsys/pkg/STATUTE-96/pdf/STATUTE-96-Pg2049.pdf (1983, accessed 18 July 2016).
3. FDA. Rare diseases: common issues in drug development, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM458485.pdf (accesssed 20 June 2016).
4. http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/sc1-pm-08-2017.html
5. Sharma A, Jacob A, Tandon M, et al. Orphan drug: development trends and strategies. *J Pharm Bioallied Sci* 2010; 2: 290-299.
6. Sasinowski F. Quantum of effectiveness evidence in FDA's approval of orphan drugs: cataloging FDA's flexibility in regulating therapies for persons with rare disorders. *Drug Inf J* 2012; 46: 238–263.
7. Sasinowski F, Panico E and Valentine J. Quantum of effectiveness evidence in FDA's approval of orphan drugs: update, July 2010 to June 2014. *Ther Innov Regul Sci* 2015; 49: 680–697.
8. https://www.congress.gov/bill/114th-congress/house-bill/6 (2015, accessed 12 April 2016).
9. Friede T, Röver C, Wandel S, et al. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods* 2017; 1: 79–91.
10. Friede T, Röver C, Wandel S, et al. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biom J*. Epub ahead of print 18 October 2016. DOI: 10.1002/bimj.201500236.
11. Spiegelhalter DJ, Abrams KR and Myles JP. Bayesian approaches to clinical trials and health-care evaluation. 1st ed. Chichester: John Wiley & Sons, 2004.
12. Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. *Clin Trials* 2010; 7: 5–18.
13. Neuenschwander B, Roychoudhury S and Schmidli H. On the use of co-data in clinical trials. *Stat Biopharm Res* 2016; 8: 345–354.
14. Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; 70: 1023–1032.
15. Gerß JWO and Köpcke W. Clinical trials and rare diseases. In: De la Paz MP and Groft SC (eds) *Advances in experimental medicine and biology*. Berlin: Springer, 2010, pp. 173–190.
16. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal* 2006; 1: 515–533.
17. Polson NG and Scott JG. On the half-cauchy prior for a global scale parameter. *Bayesian Anal* 2012; 7: 887–902.
18. Malec D. A closer look at combining data among a small number of binomial experiments. *Stat Med* 2001; 20: 1811–1824.
19. Pennello G and Thompson L. Experience with reviewing Bayesian medical device trials. *J Biopharm Stat* 2008; 18: 81–115.
20. Morita S, Thall P and Müller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008, 64: 595–602.
21. Hampson L, Whitehead J, Eleftheriou D, et al. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Stat Med* 2014; 33: 4186–4201.
22. R Core Team. R: a language and environment for statistical computing https://www.R-project.org/
23. Röver C. Bayesmeta: Bayesian random-effects meta-analysis, 2015, http://cran.r-project.org/package = bayesmeta
24. Kaye S and Choudhary A. Herpes simplex keratitis. *Prog Retin Eye Res* 2006; 25: 355–380.
25. White ML and Chodosh J. Herpes simplex virus keratitis: a treatment guideline. *Hoskins Centers Compendium of Evidence-Based Eye Care*, http://www.aao.org/clinical-statement/herpes-simplex-virus-keratitis-treatment-guideline (2014, accessed 18 January 2016).
26. Dawson CR and Togni B. Herpes simplex eye infections: clinical manifestations, pathogenesis and management. *Surv Ophthalmol* 1976; 21: 121–135.
27. Suresh PS and Tullo AB. Herpes simplex keratitis. *Indian J Ophthalmol* 1999; 47: 155–165.
28. Farooq A and Shukla D. Herpes simplex epithelial and stromal keratitis: an epidemiologic update. *Surv Ophthalmol*. 2012; 57: 448–462.
29. FDA. Zirgan approval correspondence. http://www.accessdata.fda.gov/drugsatfda_docs/nda/2009/022211s000_AdminCorres.pdf (2009, accessed 18 January 2016).

30. European Medicines Agency. Public summary of opinion on orphan designation. *Ciclosporin for the treatment of Herpes simplex virus stromal keratitis*. http://www.ema.europa.eu/docs/en_GB/document_library/Orphan_-designation/2009/10/WC500006041.pdf (2015, accessed 23 May 2016).
31. FDA. Approval letter. http://www.accessdata.fda.gov/drugsatfda_docs/appletter/2009/022211s000ltr.pdf (accesssed 18 January 2016).
32. FDA. Drug approval package: Zirgan 0.15% http://www.accessdata.fda.gov/drugsatfda_docs/nda/2009/022211_zirgan_toc.cfm (accesssed 18 January 2016).
33. Chou TY and Hong BY. Ganciclovir ophthalmic gel 0.15% for the treatment of acute herpetic keratitis: background, effectiveness, tolerability, safety, and future applications. Ther Clin Risk Manag. 2014; 10: 665–681.
34. FDA. Guidance for industry: non-inferiority clinical trials http://www.fda.gov/downloads/Drugs/.../Guidances/UCM202140.pdf (accesssed 25 May 2016).
35. Turner R, Jackson D, Wei Y, et al. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015; 34: 984–998.
36. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014; 13: 41–54.
37. Korn EL, McShane LM and Freidlin B. Statistical challenges in the evaluation of treatments for small patient populations. *Sci Transl Med*. 2013; 5: 178sr3.
38. French JA, Wang S, Warnock B, et al. Historical control monotherapy design in the treatment of epilepsy. *Epilepsia*. 2010; 51: 1936–1943.
39. FDA. Aptiom approval letter. http://www.accessdata.fda.gov/drugsatfda_docs/appletter/2015/022416Orig1s001ltr.pdf (2009, accessed 12 April 2016).
40. FDA. Aptiom label. http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/022416s001lbl.pdf (2009, accessed 12 April 2016).
41. Guyatt G, Sackett D, Taylor W, et al. Determining optimal therapy randomized trials in individual patients. *N Engl J Med* 1986; 314: 889–892.
42. Kuerner T. Essential rules and requirements for global clinical trials in rare lung diseases: a sponsor's standpoint. *Respir Investig*. 2015; 53: 2–6.
43. Heinrich MC, Joensuu H, Demetri GD, et al. Phase II, open-label study evaluating the activity of imatinib in treating life-threatening malignancies known to be associated with imatinib-sensitive tyrosine kinases. *Clin Cancer Res*. 2008; 14: 2717–2725.
44. Richter T, Nestler-Parr S, Babela R, et al. Rare Disease Terminology and Definitions – A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value Health* 2015; 18: 906–914.
45. FDA. Orphan medicines figures. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2015/04/WC500185766.pdf (2015, accessed 9 November 2016).
46. Sheffield Elicitation Framework (SHELF). http://www.tonyohagan.co.uk/shelf/ (2008, accessed 9 November 2016).
47. Varges D, Manthey H, Heinemann U, et al. Doxycycline in early CJD: a double-blinded randomised phase II and observational study. *J Neurol Neurosurg Psychiatry*. Epub ahead of print 2 November 2016. DOI: 10.1136/jnnp-2016-313541.
48. Roy ASA. Stifling new cures: the true cost of lengthy clinical drug trials. Report, Project FDA Report no. 5, 2 April 2012. New York: Manhattan Institute.
49. Warn D, Thompson S and Spiegelhalter D. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Stat Med*. 2002; 21: 1601–1623.