Statistical Science
2022, Vol. 37, No. 3, 369–379
https://doi.org/10.1214/21-STS828

Power Calculations for Replication Studies

Charlotte Micheloud and Leonhard Held

Abstract. The reproducibility crisis has led to an increasing number of replication studies being conducted. Sample sizes for replication studies are often calculated using conditional power based on the effect estimate from the original study. However, this approach is not well suited as it ignores the uncertainty of the original result. Bayesian methods are used in clinical trials to incorporate prior information into power calculations. We propose to adapt this methodology to the replication framework and promote the use of predictive instead of conditional power in the design of replication studies. Moreover, we describe how extensions of the methodology to sequential clinical trials can be tailored to replication studies. Conditional and predictive power calculated at an interim analysis are compared and we argue that predictive power is a useful tool to decide whether to stop a replication study prematurely. A recent project on the replicability of social sciences is used to illustrate the properties of the different methods.

Key words and phrases: Replication studies, conditional power, predictive power, sequential design, interim analysis, futility stopping.

1. INTRODUCTION

The replicability of research findings is essential for the credibility of science. However, the scientific world is experiencing a crisis (Begley and Ioannidis, 2015) as the replicability rate of many fields appears to be alarmingly low. As a result, large scale replication projects, where original studies are selected and replicated as closely as possible to the original procedures, have been conducted in psychology (Open Science Collaboration, 2015), social sciences (Camerer et al., 2018) and economics (Camerer et al., 2016) among others. Replication success is usually assessed using significance and p-values, compatibility of effect estimates, subjective assessments of replication teams and meta-analysis of effect estimates (e.g., in Open Science Collaboration, 2015). The statistical evaluation of replication studies is still generating much discussion and new standards are proposed (e.g., in Patil, Peng and Leek, 2016, Ly et al., 2018, Held, 2020).

Yet before a replication study is analyzed, it needs to be designed. While the conditions of the replication study are ideally identical to the original study, the replication sample size stands out as an exception and requires further consideration. Using the same sample size as in the original study may lead to a severely underpowered replication study, even if the effect $\hat{\theta}_o$ estimated in the original study is the true, unknown effect size θ (Goodman, 1992). Standard power calculations using the effect estimate from the original study as the basis for the replication study are commonly used.

A major criticism of this method is that the uncertainty accompanying this original finding is ignored and so the resulting replication study is likely to be underpowered (Anderson and Maxwell, 2017). In this paper, we propose alternatives based on predictive power and adapted from Bayesian approaches to incorporate prior knowledge to sample size calculation in clinical trials (Spiegelhalter, Abrams and Myles, 2004).

In an era where an increasing number of replication projects are being undertaken, optimal allocation of resources appears to be of particular importance. Adaptive designs are well suited for this purpose and their relevance no longer needs to be justified, particularly in clinical trials where continuing a study which should be stopped can be a matter of life or death. Stopping for futility refers to the termination of a trial when the data at interim indicate that it is unlikely to achieve statistical significance at the end of the trial (Snapinn et al., 2006). In contrast, stopping for efficacy arises when the data at interim are so convincing that there is no need to continue collecting more data. One approach for assessing efficacy and futility is called stochastic curtailment (Halperin et al., 1982), where the

Charlotte Micheloud is Ph.D. student, Department of Biostatistics, Institute of Epidemiology, Biostatistics and Prevention, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland (e-mail: charlotte.micheloud@uzh.ch). Leonhard Held is Professor, Department of Biostatistics, Institute of Epidemiology, Biostatistics and Prevention, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland (e-mail: leonhard.held@uzh.ch).

conditional power of the study, given the data so far, is calculated for a range of alternative hypotheses. Instead of conditional power, predictive power can also be used to judge if a trial should be continued (Herson, 1979). This concept has been discussed in depth in Dallow and Fina (2011) and Rufibach, Burger and Abt (2016), with an emphasis on the choice of the prior in the latter.

Lakens (2014) points out that sequential replication studies could be an alternative to fixed sample size calculations. This approach has been adopted by Camerer et al. (2018) in the *Social Sciences Replication Project* (*SSRP*), a large-scale project aiming at evaluating the replicability of social science experiments published between 2010 and 2015 in *Nature* and *Science*. A two-stage procedure was used and 21 original studies have been replicated. However, the sequential approach did not include a power calculation at interim, only allowed for a premature stopping for efficacy and did not mention any adjustment on the threshold for significance. We try to fill this gap by proposing different methods to calculate the interim power, namely the power of a replication study taking into account the data from an interim analysis. We argue that *predictive* interim power is a useful tool to guide the decision to stop replication studies where the intended effect is not present. Our framework only enables power calculation at a single interim analysis.

This paper is structured as follows: power calculations for nonsequential (Section 2) and sequential (Section 3) replication studies are presented, with a focus on comparing conditional and predictive methods. Relevant properties of these methods are then illustrated using data from the *SSRP* in Section 4. We close with some discussion in Section 5.

2. NONSEQUENTIAL REPLICATION STUDIES

Suppose a study has been conducted in order to estimate an unknown effect size θ. We consider the one-sample case throughout this paper but the results can also be generalized to the case of two samples. The study produced a positive effect estimate $\hat{\theta}_o$. In order to confirm this finding, a replication study is planned. Let us assume that the future data of the replication study are normally distributed as follows,

$$Y_1, \ldots, Y_{n_r} \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2),$$

where n_r is the replication sample size and σ the known standard deviation of one observation, assumed to be the same for original and replication study. In the *SSRP*, as well as in most replication projects, power calculations for the replication studies are based on the original effect estimate $\hat{\theta}_o$. In order to incorporate the uncertainty of $\hat{\theta}_o$ we use the following prior:

$$\theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 = \sigma^2/n_o), \tag{1}$$

centered around $\hat{\theta}_o$ and with variance inversely proportional to the original sample size n_o (Spiegelhalter, Abrams and Myles, 2004). Prior (1) may be too optimistic in practice, where original effect estimates tend to be exaggerated (Camerer et al., 2018). This issue and possible solutions are discussed in the next section.

In what follows, the different formulas resulting from the use of the prior (1) are described. This section is inspired by Section 6.5 in Spiegelhalter, Abrams and Myles (2004) where Bayesian contributions to selecting the sample size of a clinical trial are studied. We adapt this methodology to the replication framework and express the power calculation formulas in terms of unitless quantities (namely relative sample sizes and test statistics).

2.1 Methods

We differentiate between design and analysis prior, both having an impact on the power calculation (O'Hagan and Stevens, 2001), and present the different combinations of priors in Table 1. Detailed derivations of the four formulas can be found in the Supplementary Material A (Micheloud and Held, 2022a, Sections 1.1–1.4).

A point prior at $\theta = \hat{\theta}_o$ in the design corresponds to the concept of conditional power (Spiegelhalter and Freedman, 1986). In contrast, the normal design prior (1) is related to the concept of predictive power, which averages the conditional power over the possible values of the true effect according to its design prior distribution. Alternative names in the literature are assurance (O'Hagan, Stevens and Campbell, 2005), probability of study success (Wang et al., 2013) and Bayesian predictive power (Spiegelhalter, Freedman and Blackburn, 1986). Conditional and predictive power are usually accompanied by a flat analysis prior, but can also be calculated assuming that original and replication data are pooled (using the normal

TABLE 1
Methods of power calculations resulting from the different combinations of design and analysis priors

	Design	
Analysis	Point prior $\theta = \hat{\theta}_o$	Normal prior $\theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$
Flat prior	Conditional	Predictive
Normal prior $\theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$	Conditional Bayesian	Fully Bayesian

analysis prior (1)), resulting in the conditional Bayesian power and the fully Bayesian power, respectively.

In practice, publication bias and the winner's curse often lead to overestimated original effect estimates (Ioannidis, 2008, Button et al., 2013, Anderson and Maxwell, 2017). Hence, prior (1) might be over-optimistic and lead to underpowered replication studies. A simple way to correct for this over-optimism is to multiply the *design* prior mean $\hat{\theta}_o$ in (1) by a factor d between 0 and 1. The corresponding shrinkage factor $s = 1 - d$ can be chosen based on previous replication studies in the same field. This is the approach considered in the *SSRP* and we expand on this in Section 4. More advanced methods using empirical Bayes based power estimation (Jiang and Yu, 2016) and data-driven shrinkage (Pawel and Held, 2020) are not considered here.

2.1.1 *Conditional power.* Conditional power is the probability that a replication study will lead to a statistically significant conclusion at the two-sided level α, given that the alternative hypothesis is true (Spiegelhalter, Abrams and Myles, 2004, Section 2.5). In the context of a replication study, the alternative hypothesis is represented by the effect estimate $\hat{\theta}_o$ from the original study.

Let $z_{\alpha/2}$ and $\Phi[\cdot]$ respectively denote the $\alpha/2$-quantile and the cumulative distribution function of the standard normal distribution. Conditional power is

$$\text{CP} = \Phi\left[\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_{\alpha/2}\right]. \tag{2}$$

The required replication sample size n_r can be obtained by rearranging (2).

A key feature of our framework is that all power/sample size formulas are expressed without absolute effect measures. Simple mathematical rearrangements produce an expression which only depends on the original test statistic $t_o = \hat{\theta}_o/\sigma_o = \hat{\theta}_o\sqrt{n_o}/\sigma$ and the variance ratio $c = \sigma_o^2/\sigma_r^2$ which simplifies to the relative sample size $c = n_r/n_o$ and represents how much the sample size in the replication study is increased as compared to the one in the original study. Formula (2) then becomes

$$\text{CP} = \Phi[\sqrt{c}t_o + z_{\alpha/2}]. \tag{3}$$

This formula highlights an intuitive property of the conditional power: the larger the evidence in the original study (quantified by t_o) or the larger the increase in sample size compared to the original study (represented by c), the larger the conditional power of the replication study.

2.1.2 *Predictive power.* In order to incorporate the uncertainty of $\hat{\theta}_o$, the concept of predictive power is discussed (Spiegelhalter and Freedman, 1986). Its formula is

$$\text{PP} = \Phi\left[\sqrt{\frac{n_o}{n_o + n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_{\alpha/2}\right)\right]. \tag{4}$$

The predictive power (4) tends to the conditional power (3) as the original sample size n_o increases. Using the unitless quantities t_o and c, the predictive power can be rewritten as

$$\text{PP} = \Phi\left[\sqrt{\frac{c}{c+1}}t_o + \sqrt{\frac{1}{c+1}}z_{\alpha/2}\right]. \tag{5}$$

2.1.3 *Fully Bayesian and conditional Bayesian power.* So far two power calculation methods where a flat analysis prior is used have been considered. This approach corresponds to the two-trials rule in drug development, which requires "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness" (FDA, 1998, p. 3). In practice, this translates to two studies with a significant p-value and an effect in the intended direction.

An alternative approach for the analysis is to pool original and replication data. This is similar to a meta-analysis of original and replication effect estimates, as done in the *SSRP*, for example. However, in order to ensure the same evidence level as when original and replication studies are analyzed independently, the corresponding two-sided significance level $\tilde{\alpha} = \alpha^2/2$ should be used (Fisher, 1999, Gibson, 2020).

The fully Bayesian power is calculated using the prior (1) in both the design and the analysis. Using the same prior beliefs in both stages is considered as the most natural approach by some authors (e.g., in O'Hagan and Stevens, 2001). The corresponding formula is

$$\text{FBP} = \Phi\left[\sqrt{\frac{c+1}{c}}t_o + \sqrt{\frac{1}{c}}z_{\tilde{\alpha}/2}\right]. \tag{6}$$

Note that the fully Bayesian power is also a predictive power as it incorporates the uncertainty of the original effect estimate $\hat{\theta}_o$.

The last possible combination of design and analysis priors leads to the conditional Bayesian power:

$$\text{CBP} = \Phi\left[\frac{c+1}{\sqrt{c}}t_o + \sqrt{\frac{c+1}{c}}z_{\tilde{\alpha}/2}\right]. \tag{7}$$

2.2 Properties

For fixed relative sample size c and two-sided level α, all four formulas (3), (5), (6) and (7) react to an increase in original test statistic t_o with a monotone increase in power. However, the original result cannot be changed and it is more realistic to study the power when varying the relative sample size c for fixed original test statistic t_o instead. Consider two original studies with p-values 0.046 and 0.005. These p-values correspond to the original studies by Duncan, Sadanand and Davachi (2012) and Shah, Mullainathan and Shafir (2012) in the *SSRP* dataset and are used in the following for illustrative purposes. Derivations of the properties described in this section can be found in the Supplementary Material A (Micheloud and Held, 2022a, Sections 1.5–1.6).

2.2.1 *Conditional vs. predictive power.* The power obtained with predictive methods is always closer to 50% than the power obtained with conditional methods (Spiegelhalter, Abrams and Myles, 2004, Grouin et al., 2007, Dallow and Fina, 2011). In practice, power is typically larger than 50% and this implies that CP (3) is larger than PP (5); and CBP (7) is larger than FBP (6).

Furthermore, it can be shown that CP and PP are both equal to 50% if the relative sample size is

$$c = z^2_{\alpha/2}/t_o^2, \tag{8}$$

the squared $\alpha/2$-quantile of the normal distribution divided by the squared test statistic from the original study. Equation (8) implies that the larger the evidence in the original study (quantified by t_o), the smaller the relative sample size c where CP and PP curves intersect.

This can be observed in Figure 1, where the relative sample size at the intersection of the CP and PP curves is closer to zero in the replication of a convincing original study ($p_o = 0.005$, $c = 0.48$) than in the replication of a borderline original study ($p_o = 0.046$, $c = 0.96$). Likewise, FBP and CBP are crossing at a power of 50% with corresponding relative sample size

$$c = z^2_{\tilde{\alpha}/2}/t_o^2 - 1. \tag{9}$$

2.2.2 *Predictive power cannot always reach* 100%. Unlike CP (3) which always reaches 100% for a sufficiently large replication sample size, PP (5) has an asymptote at $1 - p_o/2$. This means that the more convincing the original study, the closer to 100% the PP of an infinitely large replication study is. In a sense, the original result penalizes the predictive power. However, this penalty is not very stringent, as replication of an original study with a two-sided p-value of 0.05 would still be able to reach a PP of 97.5% for a sufficiently large replication sample size. This property also applies to the FBP and can be observed in Figure 1 where the horizontal black line indicates the asymptote $1 - p_o/2$.

2.2.3 *Pooling original and replication studies.* For a borderline significant original study (e.g., $p_o = 0.046$ in Figure 1), FBP (6) and CBP (7) are, respectively, always smaller than PP (5) and CP (3). In contrast, when the original study is more convincing (e.g., $p_o = 0.005$ in Figure 1), FBP is larger than PP (respectively, CBP larger than CP) for some values of c. However, if $p_o < \tilde{\alpha}$, the level required at the end of the replication study (typically $\tilde{\alpha} = 0.00125$), FBP and CBP converge to 100% for $c \to 0$, decrease down to

$$\Phi\left[\sqrt{t_o^2 - z^2_{\tilde{\alpha}/2}}\right] \tag{10}$$

for increasing c and then increase to $1 - p_o/2$ (FBP) or 100% (CBP). A highly convincing original study will thus always have FBP and CBP very close to 100% independently of the sample size. This implies that a replication may not be required at all, a clear disadvantage of pooling original and replication studies instead of considering them independently.

3. SEQUENTIAL REPLICATION STUDIES

In Section 2, power calculations are performed before any data have been collected in the replication study. This framework is extended in this section and allows power (re)calculation at an interim analysis, after some data have been collected in the replication study already. The interim power is defined as the probability of statistical significance at the end of the replication study given the data collected so far. The incorporation of prior knowledge into interim power has been studied in Spiegelhalter, Abrams and Myles (2004), Section 6.6, and we adapt this

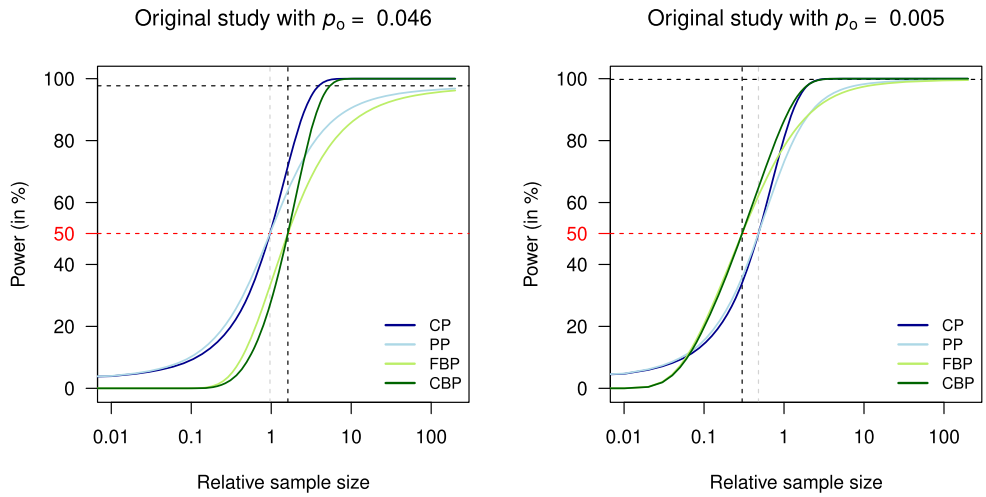

FIG. 1. *CP, PP, FBP and CBP as a function of the relative sample size c for two original studies with* $p_o = 0.046$ (*left*) *and* $p_o = 0.005$ (*right*) *at the two-sided* $\alpha = 5\%$ *level, so* $\tilde{\alpha} = 0.00125$. *The vertical grey line corresponds to the intersection of CP and PP curves as calculated in* (8), *and the vertical black line to the intersection of FBP and CBP as in* (9). *The horizontal black line indicates the asymptote* $1 - p_o/2$ *of PP and FBP.*

approach to the case where prior information refers to a single original study. Moreover, the power calculation formulas are expressed in terms of unitless quantities (relative sample sizes and test statistics) in the following. It is well known from the field of clinical trials that the maximum sample size (if the trial has not been stopped at interim) increases with the number of planned interim analyses (Matthews, 2006, Section 8.2.1). In order to maintain a given power, even one interim analysis requires a larger maximum sample size than for a trial with a fixed size and the calculation of the replication sample size should take this into account.

3.1 Methods

In addition to the point prior $\theta = \hat{\theta}_o$ and the normal prior (1), the new framework enables the specification of a flat design prior. Table 2 shows the different types of interim power calculations that are investigated in this section. Detailed derivations of the three formulas are available in the Supplementary Material A (Micheloud and Held, 2022a, Sections 2.1–2.3).

Calculating the interim power to detect the effect estimate from the original study ignores the uncertainty of the original result. This corresponds to the conditional power in Table 2. Uncertainty of the original result can be taken into account when recalculating the power at an interim analysis, turning the conditional power into a predictive power. This requires the selection of a prior distribution for the true effect, which is updated by the data collected so far in the replication study. The prior distributions discussed here are the normal prior (1) (leading to the informed predictive power) and a flat prior (leading to the predictive power). The conditional power is then averaged with respect to the posterior distribution of the true effect size, given the data already observed in the replication study. A pooled analysis of original and replication data can also be considered in this framework but is omitted here.

Let $\hat{\theta}_i$ be the effect estimate at interim and $\sigma_i^2 = \sigma^2/n_i$ the corresponding variance, with n_i the sample size at interim. The sample size that is still to be collected in the replication study is denoted by n_j and the total replication sample size is thus $n_r = n_i + n_j$. The interim power formulas can be shown to only depend on the original and interim test statistics t_o and $t_i = \hat{\theta}_i/\sigma_i$, the relative sample size $c = n_r/n_o$ and the variance ratio $f = \sigma_r^2/\sigma_i^2 = n_i/n_r$, the fraction of the replication study already completed.

3.1.1 *Conditional power at interim*. The conditional power at interim is the interim power to detect the effect $\theta = \hat{\theta}_o$. It can be expressed as

$$\text{CPi} = \Phi\left[\sqrt{c(1-f)}t_o + \sqrt{\frac{f}{1-f}}t_i + \sqrt{\frac{1}{1-f}}z_{\alpha/2}\right]. \tag{11}$$

In the particular case where no data has been collected yet in the replication study ($f = 0$), the CPi (11) reduces to the CP (3). Interim power can also be calculated to detect $\theta = \hat{\theta}_i$, this is however not recommended (Bauer and König, 2006, Kunzmann et al., 2020).

3.1.2 *Informed predictive power at interim*. The informed predictive power at interim is the predictive interim power using the design prior (1). It can be formulated as

$$\text{IPPi} = \Phi\left[\sqrt{\frac{c(1-f)}{(cf+1)(1+c)}}t_o + \sqrt{\frac{f(1+c)}{(1-f)(cf+1)}}t_i + \sqrt{\frac{cf+1}{(1+c)(1-f)}}z_{\alpha/2}\right]. \tag{12}$$

In the case of $f = 0$ (no data collected in the replication study so far), the IPPi (12) reduces to the PP (5). By considering the original result but also its uncertainty, the predictive power at interim is a compromise between considering only the original effect estimate (CPi) and ignoring the original study completely (PPi).

3.1.3 *Predictive power at interim*. The predictive power at interim is the predictive interim power using a flat design prior. In other words, the results from the original study are ignored. It is expressed as

$$\text{PPi} = \Phi\left[\sqrt{\frac{1}{1-f}}t_i + \sqrt{\frac{f}{1-f}}z_{\alpha/2}\right]. \tag{13}$$

Note that PPi (13) corresponds to FBP (6) provided that the original study in FBP formula is considered as the

TABLE 2
Methods of interim power calculations resulting from the different combinations of design and analysis priors

	Design		
Analysis	Point prior $\theta = \hat{\theta}_o$	Normal prior $\theta \sim \text{N}(\hat{\theta}_o, \sigma_o^2)$	Flat prior
Flat prior	Conditional	Informed predictive	Predictive

interim study. This illustrates the dependence of original and replication studies when a normal prior is used in the analysis.

3.2 Properties

Theoretical and specific properties of the conditional, informed predictive and predictive power at interim are discussed in this section; see Supplementary Material A (Micheloud and Held, 2022a, Sections 2.4–2.7) for additional details and derivations.

3.2.1 *Conditional vs. predictive power.* The power at interim, as compared to study start, involves two additional parameters, namely the test statistic t_i from the interim analysis and the fraction f of the replication study already conducted. It is therefore not straightforward to compare the different methods in terms of which one results in a larger power. Comparison is facilitated if certain assumptions are made. Consider any combination of a significant original result, a nonsignificant interim result and a replication sample size at least twice as large as the original sample size. This translates to $t_o > z_{1-\alpha/2}$, $t_i < z_{1-\alpha/2}$ and $c \geq 2$ in formulas (11), (12) and (13). Under these assumptions and with $f > 0.25$, the CPi is always larger than the IPPi, which is always larger than the PPi. However, one has to be careful as these conditions are sufficient, but not necessary for obtaining this order.

3.2.2 *Weights given to original and interim results.* Equations (11), (12) and (13) can be expressed as $\Phi[x]$ where x is a weighted average of t_o, t_i and $z_{\alpha/2}$ with weights w_o, w_i and w_α, say. The weights w_o and w_i depend on the relative sample size c and the fraction f of the replication study already completed.

In the CPi formula (11), an increase in c leads to a monotone increase in w_o and does not affect w_i. In other words, the weight given to the original result in the CPi becomes larger if the relative sample size c increases. Furthermore, the larger the fraction f of the replication study already completed, the less weight is given to the original result and conversely, the more weight to the interim result.

In the IPPi formula (12), an increase in f leads to a decrease in w_o and an increase in w_i. Only if the interim analysis takes place early will the original result have a greater weight than the interim result in the calculation of the IPPi.

In the PPi formula (13), no weight is given to the original result and the weight w_i given to interim results increases when f increases.

3.2.3 *A power of* 100% *cannot always be reached with the predictive methods.* Considering that an interim analysis has been conducted, n_i and t_i are fixed, and the only parameter that can vary is the sample size n_j still to be collected in the replication study. Increasing this sample size results in an increase of the relative sample size c and a decrease of the fraction f of the replication study already completed. If n_j is large enough, the CPi (11) reaches 100%. In contrast, the asymptotes of IPPi (12) and PPi (13) are penalized by the original and/or interim results. The larger the evidence in the original study and at interim (represented by t_o and t_i, respectively), the larger the asymptote of the IPPi. The asymptote of the PPi, on the other hand, is $1 - p_i/2$. This last property is explained in Dallow and Fina (2011), Section 4, and the asymptotes can be visualized in Figure 2 for an original study with $p_o = 0.005$ and two hypothetical interim results:

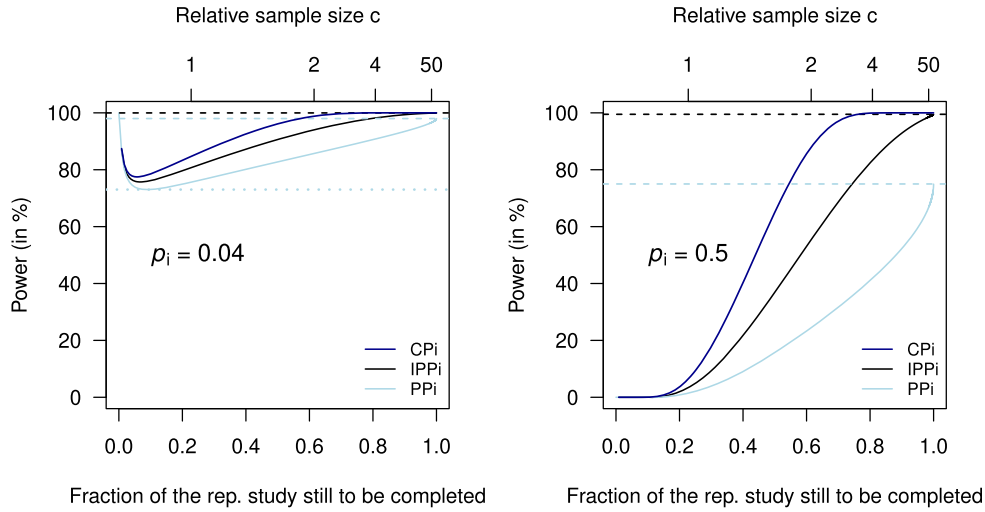

FIG. 2. *CPi, IPPi and PPi as a function of the sample size n_j still to be collected in the replication study (or equivalently, as a function of the fraction of the replication study still to be completed $(1 - f)$ and the relative sample size c) for an original study with $p_o = 0.005$ and with two hypothetical interim p-values $p_i = 0.04$ (left) and $p_i = 0.5$ (right). The two-sided level α is 0.05. Horizontal dashed lines represent the asymptotes of IPPi and PPi and the horizontal dotted line represents the minimum PPi.*

$p_i = 0.04$ and $p_i = 0.5$. On the left panel, the asymptotes of CPi, IPPi and PPi are all close to 100% as original and interim p-values are fairly small. A large increase in interim p-value hardly has an effect on the asymptote of the IPPi (from 99.98% to 99.5%, right panel) but results in a dramatic decrease of the asymptote of the PPi and remarkably, the maximum PPi achievable for a study with an interim p-value of 0.5 is only 75%.

3.2.4 *Nonmonotonicity property of power.* If the two-sided interim p-value is not significant ($p_i > \alpha$), the interim power with all three methods behaves in an expected way: it increases with increasing sample size n_j. However, this property breaks when $p_i < \alpha$. In this situation, the power assuming no additional subject to be added ($f = 1$) is 100%, declines with increasing n_j (decreasing f) and then increases. For example, the minimum predictive power at interim can be shown to be $\Phi[\sqrt{t_i^2 - z_{\alpha/2}^2}]$ which means that the PPi of any replication study with a significant interim result will never be smaller than 50%. This property can be observed in Figure 2 (left panel) where the PPi cannot be smaller than 73%. Dallow and Fina (2011) explain this characteristic as follows: "Intuitively, if the interim results are very good, any additional subject can be seen as a potential threat, able to damage the current results rather than a resource providing more power to our analysis."

4. APPLICATION

Twenty-one significant original findings were replicated in the *SSRP* and a two-stage procedure was adopted. In stage 1, the replication studies had 90% power to detect 75% of the original effect estimate. Data collection was stopped if a two-sided p-value < 0.05 and an effect in the same direction as the original effect were found. If not, data collection was continued in stage 2 to have 90% power to detect 50% of the original effect estimate for the first and second data collections pooled. The shrinkage factor s was chosen to be 0.5 as a previous replication project in the psychological field (Open Science Collaboration, 2015) found replication effect estimates on average half the size of the original effect estimates. Stages 1 and 2 can be considered as two steps of a sequential analysis, with an interim analysis in between. The analysis after stage 1 will be called the *interim* analysis while the *final* analysis will refer to the analysis based on the pooled data from stages 1 and 2.

The complete *SSRP* dataset with extended information is available at https://osf.io/pfdyw/. The effects are given as correlation coefficients, making them easily interpretable and comparable. Moreover, the application of Fisher's z transformation $z(r) = \tanh^{-1}(r)$ to the correlation coefficients justifies an asymptotic normal distribution and the standard error of the transformed coefficients becomes a function of the effective sample size $n - 3$ only, $\text{se}(z) = 1/\sqrt{n-3}$. In this dataset, original effects are always positive. A ready-to-use dataset `SSRP` can be found in the package `ReplicationSuccess`, available at https://r-forge.r-project.org/projects/replication/.

4.1 Descriptive Results

The results are displayed in Figure 3. Twelve studies were significant at interim with an effect in the correct direction but by mistake only eleven were stopped. Out of the ten studies that were continued, only two showed a significant result in the correct direction at the final analysis. The study that was wrongly continued turned out to be nonsignificant at the final analysis. The effect of publication bias is clearly seen: original effect estimates are larger than the corresponding replication effect estimates for 19 out of the 21 studies and are on average twice as large.

4.2 Power Calculations

The methods described in Sections 2 and 3 are used to calculate the power of the 21 replication studies before the onset of the study and at the interim analysis. Because our calculations are based on Fisher's z-transformed correlation coefficients, the effective sample sizes are used. The relative sample size is then $c = (n_r - 3)/(n_o - 3)$ and the fraction f of the replication study already completed $f = (n_i - 3)/(n_r - 3)$. A two-sided $\alpha = 5\%$ level is used as in the original paper, so $\tilde{\alpha} = 0.00125$ in the calculation of FBP and CBP.

4.2.1 *At the replication study start.* We computed the CP, PP, FBP and CBP of the 21 replication studies. The

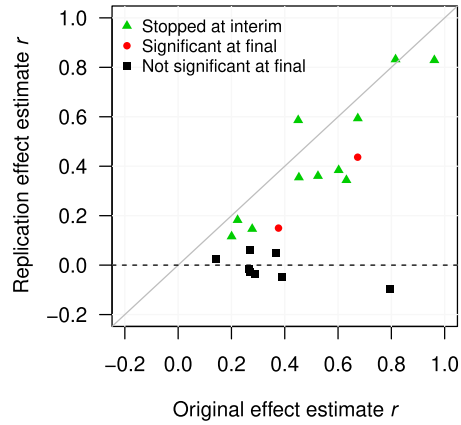

FIG. 3. *Original effect estimate vs. replication effect estimate (on the correlation scale). Replications which were not pursued in stage 2 are included with the results from stage 1. Shape and color of the point indicate whether the study was stopped due to a significant result in the correct direction at interim (green triangle), was significant in the correct direction at the final analysis (red circle) or was not significant at the final analysis (black square). The diagonal line indicates replication effect estimates equal to original effect estimates.*

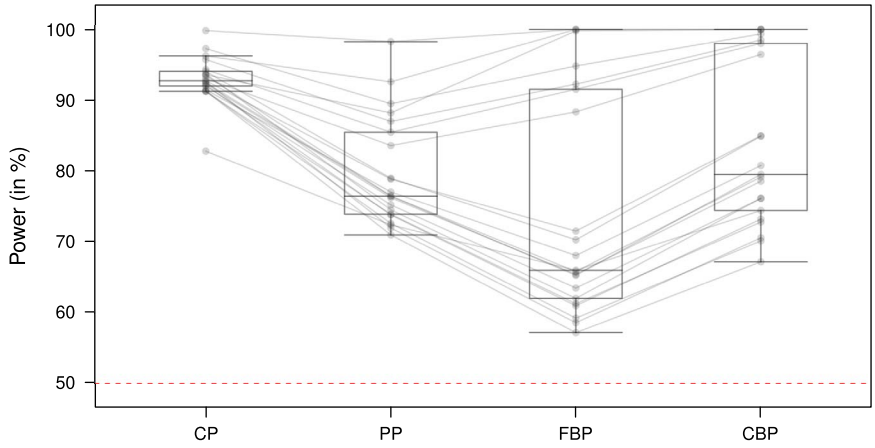

FIG. 4. *CP, PP, FBP and CBP of the* 21 *studies of the* SSRP *at level* $\alpha = 5\%$ (*so* $\tilde{\alpha} = 0.00125$ *for FBP and CBP*) *using a shrinkage factor s of* 0.25 *in the calculations. Each circle represents a study and the lines link the same studies.*

replication sample size that we considered in the calculations is the one used by the authors of the *SSRP* in stage 1, ignoring stage 2. To be consistent with the procedures of the *SSRP*, a shrinkage factor s of 0.25 was used in the calculations. Results can be found in Figure 4, where some properties discussed in Section 2.2 are illustrated. CP is larger than PP for all studies, and similarly CBP is larger than FBP as expected (see Section 2.2.1). Furthermore, it can be observed that for some studies FBP is larger than PP, while it is the opposite for some other studies. This depends on the p-value p_o from the original study and the relative sample size c as explained in Section 2.2.3. The same applies to CP and CBP but cannot be directly observed in Figure 4.

4.2.2 *At the interim analysis*. Replication studies which did not reach significance after the first data collection were continued. We have selected these studies and calculated their interim power with the different methods (see Table 3). These studies have a sample size substantially larger in the replication as compared to the original study (large c). Moreover, the interim analysis took place in the second quarter of the replication study ($0.3 \leq f \leq 0.47$) and by selection, they all have a nonsignificant interim p-value (except the study from Ackerman, Nocera and Bargh (2010) which was continued by mistake). Excluding this study, they all fulfill the sufficient conditions mentioned in Section 3.2.1 and follow the order CPi > IPPi > PPi. This also holds for the particular study with a significant interim result as the corresponding relative sample size c is large ($c = 11.62$).

The CPi is remarkably large for all studies, even for the five studies where the interim effect estimate is in the opposite direction as the original estimate as the weight given to the *significant* original result is consequent due to the large relative sample size c (see Section 3.2.2). In contrast, more weight is given to the interim as compared to the original result in the IPPi formula, making the corresponding IPPi values more sensible. If a futility bound-

TABLE 3

CPi, IPPi and PPi of the ten studies that were continued including the original, interim and replication two-sided p-values and effect estimates, the relative sample size c and the fraction f of the replication study already completed

	Original		Interim			Interim power			Replication		
Study	p_o	r_o	f	p_i	r_i	CPi	IPPi	PPi	c	p_r	r_r
Duncan	0.005	0.67	0.37	0.29	0.18	100.0	74.6	43.4	7.42	0.00001	0.44
Pyc	0.023	0.38	0.43	0.09	0.15	100.0	85.3	71.0	9.18	0.009	0.15
Ackerman	0.048	0.27	0.43	0.02	0.14	100.0	95.0	90.3	11.69	0.125	0.06
Rand	0.009	0.14	0.47	0.37	0.03	99.8	51.9	27.0	6.27	0.234	0.03
Ramirez	0.000008	0.79	0.30	0.72	−0.08	100.0	61.4	4.2	4.47	0.390	−0.10
Gervais	0.029	0.29	0.42	0.41	−0.05	97.5	1.9	0.3	9.78	0.415	−0.04
Lee	0.013	0.39	0.42	0.45	−0.07	97.7	3.1	0.4	7.65	0.435	−0.05
Sparrow	0.002	0.37	0.44	0.27	0.11	99.7	74.1	40.1	3.50	0.451	0.05
Kidd	0.012	0.27	0.40	0.27	−0.07	98.9	1.6	0.1	8.57	0.467	−0.03
Shah	0.046	0.27	0.45	0.15	−0.09	87.0	0.1	0.0	11.62	0.710	−0.02

ary between 10% and 30% had been used (as in DeMets (2006)), four out of the eight studies which failed to replicate at the final analysis would have been stopped at interim based on the IPPi values. Surprisingly, the replication study of Ramirez and Beilock (2011) presents a relatively large IPPi (61.4%) although the interim result goes in the opposite direction as the original result. This is due to the very small original p-value. The PPi of the same study is considerably smaller (4.2%) since the original result does not influence the power with this method. Furthermore, six out of eight studies which failed to replicate at the final analysis would have been stopped at interim if futility stopping had been decided based on a PPi of less than 30%. Significant interim results lead to large PPi values (see Section 3.2.4), and that can observed for the study that was incorrectly continued.

5. DISCUSSION

Conditional power calculations appear to be the norm in most replication projects. In this paper, we have drawn attention to notable shortcomings of this approach and outlined the rationale and properties of predictive power. We encourage researchers to abandon conditional methods in favor of predictive methods which make a better use of the original study and its uncertainty.

Furthermore, as many replications are being conducted and only a fraction confirms the original result, we argue for the necessity of sequentially analyzing the results. With this in mind, we encourage the initiative from Camerer et al. (2018) to terminate some replication studies prematurely based on an interim analysis. However, their approach only enables efficacy stopping. We propose to use interim power to judge if a replication study should be stopped for futility. Interim analyses can help to save time and resources but also raise new questions with regard to the choice of prior distributions. We have shown using studies from the *SSRP* that different design priors lead to very different power values and by extension to different decisions. Conditioning the power calculations at interim on the original results is even more unreasonable than at the study start and leads to very large power values given a significant original result, even if interim results suggest evidence in the opposite direction. We recommend the use of IPPi and PPi to make futility decisions. A 30% futility boundary is sometimes employed in clinical trials and has proved to be reasonable in the *SSRP*. Efficacy stopping based on interim power is known to inflate the type-I error rate (Jennison and Turnbull, 1999, Chapter 10). We only consider futility stopping as this issue does not apply here (Lachin, 2005).

Some limitations should be noted. First, the paper discusses power calculations before the onset of the study and at an interim analysis separately. However, the planned interim analysis has an impact on power at study start and sample size adjustments are necessary (Wassmer and Brannath, 2016, Section 2.1.2). This is nevertheless rarely done in current replication projects such as *SSRP*. Second, while the ICH E9 'Statistical Principles for Clinical Trials' (ICH E9 Expert Working Group, 1999) recommends blinded interim results, our data at interim are assumed to be unblinded. This is not a problem for the one-sample case but becomes an issue when we want to compare two groups. Such a situation would require an Independent Data Monitoring Committee to prevent the replication study from being biased (Kieser and Friede, 2003). Third, the assumption of normally distributed observations is made.

Further research will focus on extending this framework to multiple interim analyses in a replication study and to sequentially conducted replication studies. It will also be of interest to apply the concept of interim power discussed in Section 3 to the reverse-Bayes assessment of replication success (Held, 2020).

SOFTWARE

Software for these power calculations can be found in the R-package `ReplicationSuccess`, available at https://r-forge.r-project.org/projects/replication/. An example of the usage of this package is given in the Supplementary Material B (Micheloud and Held, 2022b).

ACKNOWLEDGMENTS

We thank Samuel Pawel, Małgorzata Roos and Lawrence L. Kupper for helpful comments and suggestions on this manuscript. We also would like to thank the referees whose comments helped to improve and clarify the manuscript.

FUNDING

This work was funded by the Swiss National Science Foundation (project 189295).

SUPPLEMENTARY MATERIAL

Supplement to "Power Calculations for Replication Studies" (DOI: 10.1214/21-STS828SUPP; .zip). Derivation of the formulas in Sections 2 and 3. Example of the usage of the R-package `ReplicationSuccess`.

REFERENCES

ACKERMAN, J. M., NOCERA, C. C. and BARGH, J. A. (2010). Incidental haptic sensations influence social judgments and decisions. *Science* **328** 1712–1715. https://doi.org/10.1126/science.1189993

ANDERSON, S. F. and MAXWELL, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivar. Behav. Res.* **52** 305–324. https://doi.org/10.1080/00273171.2017.1289361

BAUER, P. and KÖNIG, F. (2006). The reassessment of trial perspectives from interim data–a critical view. *Stat. Med.* **25** 23–36. https://doi.org/10.1002/sim.2180

BEGLEY, C. G. and IOANNIDIS, J. P. A. (2015). Reproducibility in science. *Circ. Res.* **116** 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

BUTTON, K. S., IOANNIDIS, J. P., MOKRYSZ, C., NOSEK, B. A., FLINT, J., ROBINSON, E. S. and MUNAFÒ, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14** 365. https://doi.org/10.1038/nrn3475

CAMERER, C. F., DREBER, A., FORSELL, E., HO, T.-H., HUBER, J., JOHANNESSON, M., KIRCHLER, M., ALMENBERG, J., ALTMEJD, A. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* **351** 1433–1436. https://doi.org/10.1126/science.aaf0918

CAMERER, C. F., DREBER, A., HOLZMEISTER, F., HO, T.-H., HUBER, J., JOHANNESSON, M., KIRCHLER, M., NAVE, G., NOSEK, B. A. et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2** 637–644. https://doi.org/10.1038/s41562-018-0399-z

DALLOW, N. and FINA, P. (2011). The perils with the misuse of predictive power. *Pharm. Stat.* **10** 311–317. https://doi.org/10.1002/pst.467

DEMETS, D. L. (2006). Futility approaches to interim monitoring by data monitoring committees. *Clin. Trials* **3** 522–529. https://doi.org/10.1177/1740774506073115

DUNCAN, K., SADANAND, A. and DAVACHI, L. (2012). Memory's penumbra: Episodic memory decisions induce lingering mnemonic biases. *Science* **337** 485–487. https://doi.org/10.1126/science.1221936

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. Available at www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products.

FISHER, L. D. (1999). One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. *Drug Inf. J.* **33** 265–271. https://doi.org/10.1177/009286159903300130

GIBSON, E. W. (2020). The role of p-values in judging the strength of evidence and realistic replication expectations. *Stat. Biopharm. Res.* **13** 6–18. https://doi.org/10.1080/19466315.2020.1724560

GOODMAN, S. N. (1992). A comment on replication, P-values and evidence. *Stat. Med.* **11** 875–879. https://doi.org/10.1002/sim.4780110705

GROUIN, J.-M., COSTE, M., BUNOUF, P. and LECOUTRE, B. (2007). Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations. *Stat. Med.* **26** 4914–4924. https://doi.org/10.1002/sim.2958

HALPERIN, M., LAN, K. K. G., WARE, J. H., JOHNSON, N. J. and DEMETS, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Control. Clin. Trials* **3** 311–323. https://doi.org/10.1016/0197-2456(82)90022-8

HELD, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *J. R. Stat. Soc., A* **183** 431–469. https://doi.org/10.1111/rssa.12493

HERSON, J. (1979). Predictive probability early termination plans for phase II clinical trials. *Biometrics* **35** 775–783.

ICH E9 EXPERT WORKING GROUP (1999). Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Stat. Med.* **18** 1905–1942.

IOANNIDIS, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology* **19** 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

JENNISON, C. and TURNBULL, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials*. CRC Press/CRC, Boca Raton. https://doi.org/10.1201/9780367805326

JIANG, W. and YU, W. (2016). Power estimation and sample size determination for replication studies of genome-wide association studies. *BMC Genomics* **17** 19–32. https://doi.org/10.1186/s12864-017-3482-3

KIESER, M. and FRIEDE, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Stat. Med.* **22** 3571–3581. https://doi.org/10.1002/sim.1585

KUNZMANN, K., GRAYLING, M. J., LEE, K. M., ROBERTSON, D. S., RUFIBACH, K. and WASON, J. M. S. (2020). Conditional power and friends: The why and how of (un)planned, unblinded sample size recalculations in confirmatory trials. Technical Report. Available at https://arxiv.org/abs/2010.06567.

LACHIN, J. M. (2005). A review of methods for futility stopping based on conditional power. *Stat. Med.* **24** 2747–2764. https://doi.org/10.1002/sim.2151

LAKENS, D. (2014). Performing high-powered studies efficiently with sequential analyses. *Eur. J. Soc. Psychol.* **44** 701–710. https://doi.org/10.1002/ejsp.2023

LY, A., ETZ, A., MARSMAN, M. and WAGENMAKERS, E.-J. (2018). Replication Bayes factors from evidence updating. *Behav. Res. Methods* **51** 2498–2508. https://doi.org/10.3758/s13428-018-1092-x

MATTHEWS, J. N. (2006). *Introduction to Randomized Controlled Clinical Trials*. CRC Press/CRC, Boca Raton. https://doi.org/10.1201/9781420011302

MICHELOUD, C. and HELD, L. (2022a). Supplement A to "Power Calculations for Replication Studies." https://doi.org/10.1214/21-STS828SUPP

MICHELOUD, C. and HELD, L. (2022b). Supplement B to "Power Calculations for Replication Studies." https://doi.org/10.1214/21-STS828SUPP

O'HAGAN, A. and STEVENS, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med. Decis. Mak.* **21** 219–230. https://doi.org/10.1177/0272989X0102100307

O'HAGAN, A., STEVENS, J. W. and CAMPBELL, M. J. (2005). Assurance in clinical trial design. *Pharm. Stat.* **4** 187–201. https://doi.org/10.1002/pst.175

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349** aac4716. https://doi.org/10.1126/science.aac4716

PATIL, P., PENG, R. D. and LEEK, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* **11** 539–544. https://doi.org/10.1177/1745691616646366

PAWEL, S. and HELD, L. (2020). Probabilistic forecasting of replication studies. *PLoS ONE* **15** e0231416. https://doi.org/10.1371/journal.pone.0231416

RAMIREZ, G. and BEILOCK, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science* **331** 211–213. https://doi.org/10.1126/science.1199427

RUFIBACH, K., BURGER, H. U. and ABT, M. (2016). Bayesian predictive power: Choice of prior and some recommendations for its use as probability of success in drug development. *Pharm. Stat.* **15** 438–446. https://doi.org/10.1002/pst.1764

SHAH, A. K., MULLAINATHAN, S. and SHAFIR, E. (2012). Some consequences of having too little. *Science* **338** 682–685. https://doi.org/10.1126/science.1222426

SNAPINN, S., CHEN, M.-G., JIANG, Q. and KOUTSOUKOS, T. (2006). Assessment of futility in clinical trials. *Pharm. Stat.* **5** 273–281. https://doi.org/10.1002/pst.216

SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York. https://doi.org/10.1002/0470092602

SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat. Med.* **5** 1–13. https://doi.org/10.1002/sim.4780050103

SPIEGELHALTER, D. J., FREEDMAN, L. S. and BLACKBURN, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Control. Clin. Trials* **7** 8–17. https://doi.org/10.1016/0197-2456(86)90003-6

WANG, Y., FU, H., KULKARNI, P. and KAISER, C. (2013). Evaluating and utilizing probability of study success in clinical development. *Clin. Trials* **10** 407–413. https://doi.org/10.1177/1740774513478229

WASSMER, G. and BRANNATH, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer, Berlin. https://doi.org/10.1007/978-3-319-32562-0