# Predictive probability of success in clinical drug development

Mauro Gasparini[1], Lilla Di Scala[2], Frank Bretz[3], Amy Racine-Poon[3]

## Abstract

Predictive probability of success is a (subjective) Bayesian evaluation of the probability of a future successful event in a given state of information. In the context of pharmaceutical clinical drug development, successful events relate to the accrual of positive evidence on the therapy which is being developed, like demonstration of superior efficacy or ascertainment of safety. Positive evidence will usually be obtained via standard frequentist tools, according to the regulations imposed in the world of pharmaceutical development. Within a single trial, predictive probability of success can be identified with expected power, i.e. the evaluation of the success probability of the trial. Success means, for example, obtaining a significant result of a standard superiority test. Across trials, predictive probability of success can be the probability of a successful completion of an entire part of clinical development, for example a successful phase III development in the presence of phase II data. Calculations of predictive probability of success in the presence of normal data with known variance will be illustrated, both for within-trial and across-trial predictions.

*Keywords: Interim analysis; interim power; equivalence trial; expected power; predictive distribution.*

***DOI:* 10.2427/8760**

## 1 INTRODUCTION

*Predictive probability of success* (PPS from now on), is a subjective Bayesian evaluation of the probability of a successful event in a given state of information. In pharmaceutical clinical drug development, successful events relate to the accrual of positive evidence related to the therapy which is being developed. Clinical drug development is thought here as a series of logically connected clinical trials aimed at building evidence in favor of an experimental therapy.

The world of clinical trials aimed at submission of a new drug application is highly regulated. Standard practice often requires the *sponsor* (usually a pharmaceutical company) to analyse its results according to the usual prescriptions of classical statistics. In particular, Neyman-Pearson type tests, $p$-values and confidence intervals are standard tools. They provide the proper language by which a sponsor communicates to Regulatory Agencies (RAs from now on) and agrees on

[1] ***Correspondig Author***, Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy. *e-mail*: mauro.gasparini@polito.it

[2] Biostatistics Oncology, Hoffmann-La Roche, Basel, Switzerland. *e-mail*: lilla.di_scala@roche.com

[3] Statistical Methodology, Novartis Pharma AG, Basel, Switzerland. *e-mail*: frank.bretz@novartis.com

[3] Modeling & Simulation, Novartis Pharma AG, Basel, Switzerland. *e-mail*: amy.racine@novartis.com

**Table 1.** Success predictions under different states of information

| Success prediction | states of information |
|---|---|
| about next trial | before trial begins |
| about current trial | at interim analysis |
| about phase III | at the end of phase II |

planning relevant analyses. The general picture is therefore dominated by frequentist methods, although there have been recent signs of interests on Bayesian methods by several RAs.

On the other hand, the sponsor is also a business operator and owns a rich background of scientific and practical experience with similar trials. It is therefore reasonable for a sponsor planning, predicting and conducting trials to make use of all relevant business and scientific information. One available statistical technology devoted to that purpose is the Bayesian reasoning, which has two well known advantages. Firstly, it can model appropriately the updating and the accumulation of knowledge and scientific experience from trial to trial in different states of information. Secondly, the Bayesian mechanism can accomodate economical considerations into statistical practice. The usual critique to Bayesianism, subjectivity, is not relevant here, since what is being advocated is conducting trials according to the generally agreed frequentist rules, but using all relevant information for decision making, regarding business plans and the prediction of costs and benefits.

PPS is a quantification of the probability of an event that characterizes successful clinical development as it unrolls. Predictions can be made at different states of information, depending on which point of clinical development has been reached. Table 1 contains some examples, explained in more detail in the following sections. We can distinguish PPS relative to a single trial (within-trial success predictions) and PPS relative to an entire stage of clinical development, like the sequence of two or more trials (across-trial success predictions). The former analysis has been more formally developed in the literature, especially in [1], where the term “assurance” is used instead of PPS, and it overlaps with the rich literature on Bayesian monitoring of clinical trials, see for example [2]. Across-trial predictions are less common, but examples have already appeared in the literature, especially regarding the phase II to phase III transition, where decisions have to be made and several kinds of business and scientific risks have to be evaluated. See for example [3] and the very detailed [4].

Section 2 is a review of within-trial PPS in the presence of normal observations for a variety of designs typically encountered in clinical trials. In Section 3, consideration of interim analysis is added and the relationship between PPS and the literature on Bayesian monitoring of interim analysis is reviewed. In Section 4 the general concept of across-trial PPS is formalized and illustrated within the context of transition from phase II to phase III.

## 2 WITHIN-TRIAL SUCCESS PREDICTIONS

The Bayesian setup considered in this section is the following: in a clinical trial, information is collected about a treatment effect parameter $\delta$, which is also assigned a prior probability law describing the uncertainty of the sponsor. Success of a trial is defined as rejecting a given null hypothesis $H_0$ involving $\delta$. From a Bayesian viewpoint, the usual power is the *conditional* probability of success, a random variable function of the unknown $\delta$.

PPS is then, in this case, the *unconditional*, or *expected* power, i.e. the expectation of the random variable power with respect to the current distribution of $\delta$. In the literature, PPS has also been called *predictive power*, to emphasize that calculations are done with respect to the predictive (marginal) distribution of the clinical data.

PPS as expected power was first introduced in [5] and [6], where it is exemplified in the context of the interim analysis evaluation of a parallel design trial with Bernoulli outcomes. Bernoulli outcomes were also studied by [7]. The concept is expanded in books [2] and [8] and it has also been reviewed in [9]. The normal, parallel design, case was considered by [10] and more thoroughly discussed by [11] and [1]. Critical positions are contained in [12], which takes a more fundamentalist Bayesian viewpoint, and in the recent [13], which illustrates some practical delicate issues.

## 2.1 Superiority trials with parallel design

A *superiority trial* with parallel groups is a basic and common design in clinical trials, and it is intended to show superiority of a treatment against an active control (for example, the best available treatment) or placebo. It is often the case that a very small number of superiority trials with a sufficiently large number of patients is used by the sponsor for registration by the RAs.

Let two samples of sizes $n$ (for treatment, TRT) and $m$ (for control, CTR) be independent normal random samples with unknown means $\mu_{\text{TRT}}$ and $\mu_{\text{CTR}}$, respectively, and common known variance $\sigma^2$. The unknown means are to be compared in terms of the sample mean difference $D = \bar{X}_{\text{TRT}} - \bar{X}_{\text{CTR}}$ which is normally distributed with unknown mean treatment effect $\delta = \mu_{\text{TRT}} - \mu_{\text{CTR}}$ and known variance $s^2 = \sigma^2(\frac{1}{n} + \frac{1}{m})$. The resulting normal distribution of $D$, conditionally on $\delta$, is written, in short, $D|\delta \sim \mathcal{N}\big(\delta, s^2\big)$.

Suppose the larger the better, for the sake of definiteness, that is, the higher the difference the more beneficial the treatment appears. For example, such may be the case if the primary endpoint were the CD4+ cell count per millilitre of blood in patients infected by HIV, or the FEV (forced expiratory volume) in a respiratory trial. According to standard practice, the relevant statistical technique is a test of the null hypothesis $H_0 : \delta \leq \delta_0$, where $\delta_0$, often equal to zero, is a value of clinical indifference. $H_0$ is going to be rejected if $D > \delta_0 + z_\alpha s$, where the *cut-off point* notation has been adopted for $z_\alpha$, which is defined by $\mathrm{P}(Z > z_\alpha) = \alpha$, with $Z$ standard normal.

Bayesian calculations simplify if $\delta$ is taken to be normally distributed, with mean $\theta$ and variance $\tau^2$: in short, $\delta \sim \mathcal{N}\big(\theta, \tau^2\big)$. This is the well known framework of a conjugate Bayesian analysis and standard calculations show that, marginally,

$$D \sim \mathcal{N}(\theta, \tau^2 + s^2).$$

It is then easy to compute PPS in this first case of a superiority trial at the planning stage as

$$\text{PPS} = \mathrm{P}(D > \delta_0 + z_\alpha s) = 1 - F_D(\delta_0 + z_\alpha s|\theta, \tau^2 + s^2), \tag{1}$$

where, from now on, $F_X(x|m, v^2)$ indicates the normal distribution function of a random variable $X$ computed at $x$, with $X$ normally distributed with mean $m$ and variance $v^2$.

## 2.2 First example: a superiority clinical trial

Suppose a study is being designed for a specific compound $TRT$, in which a well established clinical endpoint $Y$, measured in minutes, is available. Suppose the between-patient variation of $Y$ is approximately 50 minutes, while a treatment difference of 10 minutes would imply that $TRT$ is considered as clinically better than the standard.

The aim is to design an appropriate one-sided superiority trial with parallel groups to test the null hypothesis $H_0 : \delta \leq 0$ that $TRT$ is not superior to $CTR$. The standard deviation is $\sigma = 50$, and suppose the level of the test is $\alpha = 0.05$. Let the alternative value $\delta_A = 10$ be the basis for standard power and sample size calculations. Imposing equal sample sizes $n = m$ for both $TRT$ and $CTR$ groups, in order to achieve at least $1 - \beta = 0.95$ power at $\delta_A = 10$, according to the usual formula,

$$n = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{\delta_A^2} = 541$$

patients per arm would be required. $H_0$ will then be rejected and the treatment demonstrated superior to control if, after observing 541 patients per arm, the following event will be observed: $D > \delta_0 + z_\alpha s = 0 + 1.645 \times 50 \times 2/541 = 0.304$.

With respect to the elicitation of the prior, suppose now the sponsor is fairly confident of achieving the alternative value $\delta_A = 10$. In other words, a treatment effect $\delta_A = 10$ is not only the smallest true difference worth detecting, but it also equals a value which the sponsor considers most likely. On the basis of these considerations, the prior mean for $\delta$ is taken to be $\theta = 10$. As for the prior variance, assume the sponsor has some difficulty in choosing a particular value (in these applied setups, it is easier to elicit location parameters from laypersons, whereas dispersion parameters are harder to understand and identify). In the lack of a better understanding, a prior variance is chosen so that the prior probability that $\delta < 0$ equals 0.05. In other words, the sponsor believes in the alternative, but is willing to allow for a 5% chance that this belief is totally wrong. The uncertainty about the treatment difference $\delta$ is finally expressed as a normal prior with mean $\theta = 10$ and standard deviation $\tau = 6.08$. Equation (1)) yields PPS=0.77. Notice that this value deviates markedly from the power at $\delta_A$, which is 0.95. The latter large value may convey a false sense of security and induce the naive sponsor to be fairly optimistic about the trial success. Such optimism is not granted, since a power statement is a hypothetical statement about the chance of success given the alternative, and not a balanced evaluation of the possibility of success. PPS, or expected probability, gives instead a more realistic perspective by balancing high expectations from the new treatment and a more skeptical attitude. A naïve user of the power approach considers $\delta_A$ as fixed, while neglecting the uncertainty about the magnitude of the mean difference. Such a dichotomous approach ('either $\delta_0$ or $\delta_A$ is true') might lead to excessively high expectations regarding the results of experimentation.

### 2.3 Equivalence and noninferiority trials

An important problem in clinical trials is to establish either *equivalence* or *noninferiority* of two treatments. In an equivalence trial the traditional roles of the null and the alternative hypothesis are switched, since the sponsor wishes to "prove" similarity, for example, of two means. The standard solution is a decision rule to accept equivalence if a confidence interval for the difference of effects is fully contained within limits which are considered to have the same therapeutic effect. The limits are often symmetric around the null effect. See [14] or a more recent discussion in [15]. An equivalence trial with one of the equivalence limits set to $+\infty$ or $-\infty$ is also called a noninferiority trial.

Let then $L$ and $U$, possibly $\pm\infty$, be the lower and upper equivalence limits agreed with RAs. Two-sided equivalence at level $\alpha$ is claimed if the $(1-2\alpha)$ level confidence interval for $\mu_{\text{TRT}} - \mu_{\text{CTR}}$ is completely contained within the interval $[L, U]$; i.e. if $\{L \leq D - z_\alpha s \cap D + z_\alpha s \leq U\}$ or, equivalently, if $\{L + z_\alpha s \leq D \leq U - z_\alpha s\}$. PPS is the probability of success, evaluated before

**Table 2.** Cell mean model for a standard 2x2 cross-over trial

| | Period | |
|---|---|---|
| Sequence group | 1 | 2 |
| TREAT-CONTR | $\mu+\pi+\delta/2$ | $\mu-\pi-\delta/2$ |
| CONTR-TREAT | $\mu+\pi-\delta/2$ | $\mu-\pi+\delta/2$ |

trial begins, that is, the probability of eventually claiming equivalence:

$$\begin{aligned}\text{PPS} &= \text{P}(L + z_\alpha s \le D \le U - z_\alpha s)\\ &= F_D(U - z_\alpha s|\theta, \tau^2 + s^2) - F_D(L + z_\alpha s|\theta, \tau^2 + s^2)\end{aligned}$$

if $U - z_\alpha s > L + z_\alpha s$, 0 otherwise.

## 2.4 Cross-over trials

In clinical trials, a sometimes viable alternative to the parallel design is the *cross-over design*. In the two-treatment two-period cross-over design, each patient receives both treatments in a randomized sequence (treatment-control or control-treatment) in two different periods, separated by a wash-out phase. Superiority, noninferiority and equivalence trials may all be conducted with a cross-over design, which is used when there is hope to lower the variability of the statistics by making each subject to be a control of oneself. For radical treatments like surgical, cross-over trials are obviously not possible. For a thorough discussion of cross-over trials, see for example [16].

Consider the classical model proposed by [17] about the two-treatment two-period cross-over design for a clinical trial. In its simplest form, without carry-over effect and with the same number $n$ of observations in each group, the observation from the $k$-th period of the $j$-th patient, to whom sequence $i$ is administered, can be written

$$Y_{ijk} = \mu + \pi_k + \tau_l + \eta_{ij} + \epsilon_{ijk} \quad i = 1, 2,\ j = 1, \dots, n,\ k = 1, 2,\ l = 1, 2$$

where $\mu, \pi_k, \tau_l$ are overall mean, period and treatment effects, $\eta_{ij}$ are random independent patient effects with mean 0 and variance $\sigma^2_\eta$ - the interpatient variance - and $\epsilon_{ijk}$ are independent normally distributed error effects with mean 0 and intrapatient variance $\sigma^2_\epsilon$, assumed to be known. Reparameterizing with $\pi = \pi_1 = -\pi_2$ and $\delta = 2\tau_1 = -2\tau_2$ the cell mean model contained in Table 2 can be obtained.
The least square estimator of the treatment effect $\delta$ is

$$D_c = \frac{1}{2}(\bar{y}_{1.1} - \bar{y}_{1.2} - \bar{y}_{2.1} + \bar{y}_{2.2})$$

where $\bar{y}_{i.k} = \sum_j y_{ijk}/n$. Since $y_{ij1} - y_{ij2}$ are conditionally independent normal random variables with variance $2\sigma^2_\epsilon$, it can be seen that $D_c|\delta \sim \mathcal{N}(\delta, s^2_c)$ where $s^2_c = \sigma^2_\epsilon/n$ is the intrapatient variance, divided by the number of patients per sequence group.

For superiority trials, the calculation of PPS is formally the same as for the parallel group design, provided that the correct distribution above is used for the treatment effect estimate:

$$\text{PPS} = 1 - F_{D_c}(\delta_0 + z_\alpha s_c|\theta, \tau^2 + s^2_c) \qquad (2)$$

**Table 3.** PPS corresponding to three different prior distributions

| type of prior | prior mean $\theta$ | PPS |
|---|---|---|
| skeptical prior | 0 | 0.40 |
| conventional prior | 1.5 | 0.71 |
| optimistic prior | 3 | 0.92 |

Similarly, for equivalence trials, if $L$ and $U$ are the lower and upper equivalence bounds, PPS becomes

$$\mathrm{PPS} = F_{D_c}(U - z_\alpha s_c|\theta, \tau^2 + s_c^2) - F_{D_c}(L + z_\alpha s_c|\theta, \tau^2 + s_c^2) \tag{3}$$

if $U - z_\alpha s_c > L + z_\alpha s_c$, 0 otherwise.

### 2.5 Second example: a cross-over trial with a variety of priors

Consider a new therapy meant to reduce diastolic blood pressure (DBP). Patients studied in related clinical trials usually have DBP between 95 and 115 mm/hg. Suppose the standard treatment reduces DBP from baseline by 7 mm/hg after eight weeks on the average (placebo itself reduces it by about 2 mm/hg) with an intra-patient standard deviation of about 2 mm/hg. The sponsor hopes the new therapy will lower average DBP by an extra 3 mm/hg.

The new therapy will be compared against standard therapy in a cross-over superiority trial involving $n = 200$ patients per sequence. The value $\delta_0 = 0$ represents clinical indifference and $\delta_A = 3$ is the clinically relevant alternative. Suppose the choice $\alpha = 0.01$ is declared in the protocol.

For Bayesian calculations, three priors for $\delta$ are considered, as done systematically in [18]: a skeptical one, centered around $\theta = 0$, an optimistic one, centered around $\theta = 3$, and a "middle of the road" conventional third choice, centered around $\theta = 1.5$. All three priors are given the same standard deviations, obtained by letting the prior probability that $\delta > 3$ under the skeptical prior equal 0.05 or, symmetrically, by letting the prior probability that $\delta < 0$ under the optimistic prior equal 0.05. The resulting prior standard deviation of $\delta$ is $\tau = 3/z_{0.05} = 1.82$. PPS calculations (equation (2)) corresponding to the three priors are reported in Table 3.

The point of this example is that when the sponsor is faced with a Bayesian model and does not want to commit to a single prior, the recommendation is to use "a community of priors", from pessimistic to optimistic ones. The different priors can reflect the different points of view that can arise, internally or externally to the sponsor.

## 3 WITHIN-TRIAL SUCCESS PREDICTIONS WITH INTERIM ANALYSIS

In clinical trials, *group sequential* analysis implies one or more *interim analyses*, or preplanned "looks" at the partial data, often included in a parallel design. In agreement with the RAs, the sponsor plans for interim analyses in particularly sensitive situations when, from an ethical and/or economical point of view, it is important to allow for early termination of a trial providing particularly convincing results, in the positive or the negative direction. See for example [19] and bibliography therein.

Berry in [20] promotes the conditional point of view in inferential statistics and uses interim analysis in clinical trials as a particularly stringent example. In particular, he favors a complete Bayesian approach to interim analysis, as also do the authors of [21]. According to the hybrid approach taken in this paper instead, it is assumed that an interim analysis is officially conducted according to standard practice, but that it is also important for the sponsor to have an evaluation of PPS which takes into account all the information available at the interim analysis stage.

When an interim analysis is included in a clinical trial protocol, PPS calculations can then be done either at the planning stage, before trial begins, and at the interim evaluation stage, when interim results become available and the probability of trial success has to be revised. For the sake of simplicity, only one interim look is contemplated in the following sections.

### 3.1 Planning stage

Suppose one interim look is planned for the time a specified fraction of observations has been sampled. At the planning stage, i.e. before the trial begins, the calculation of PPS has to take into account the possibility of rejecting the null hypothesis either at the interim or at the final stage.

Let $n_1$ be the planned number of treatment observations at the interim stage and $m_1$ the number of control observations and suppose that, up to negligible rounding errors, $n_1/n = m_1/m$. Let $D_1$ be the sample mean difference at the interim stage and $D_2$ be the sample mean difference of the observations taken after interim analysis so that, at the end of the trial,

$$D = \frac{\sum_1^{n_1} X_{\text{TRT}} + \sum_{n_1+1}^{n} X_{\text{TRT}}}{n} - \frac{\sum_1^{m_1} X_{\text{CTR}} + \sum_{m_1+1}^{m} X_{\text{CTR}}}{m} = \frac{n_1 D_1 + n_2 D_2}{n}.$$

The statistics $D_1$ and $D_2$ are normally distributed and conditionally independent given $\delta$, with standard errors

$$s_i = \sqrt{\sigma^2(\frac{1}{n_i} + \frac{1}{m_i})}, \quad i = 1, 2.$$

Their joint marginal distribution, which is needed in order to calculate expected power, is bivariate normal with means

$$\mathrm{E}\left(D_1\right) = \mathrm{E}\left(D_2\right) = \mathrm{E}\left(\mathrm{E}\left(D_2|\delta\right)\right) = \theta,$$

variances

$$\begin{aligned}\mathrm{Var}\left(D_1\right) &= \mathrm{Var}\left(\mathrm{E}\left(D_1|\delta\right)\right) + \mathrm{E}\left(\mathrm{Var}\left(D_1|\delta\right)\right) = \mathrm{Var}\left(\delta\right) + \mathrm{E}\left(s_1^2\right) = \tau^2 + s_1^2,\\ \mathrm{Var}\left(D_2\right) &= \mathrm{Var}\left(\mathrm{E}\left(D_2|\delta\right)\right) + \mathrm{E}\left(\mathrm{Var}\left(D_2|\delta\right)\right) = \mathrm{Var}\left(\delta\right) + \mathrm{E}\left(s_2^2\right) = \tau^2 + s_2^2\end{aligned}$$

and covariance $\mathrm{Cov}\left(D_1, D_2\right) = \mathrm{E}\left(\mathrm{E}\left(D_1 D_2|\delta\right)\right) - \mathrm{E}\left(D_1\right)\mathrm{E}\left(D_2\right) = \mathrm{E}\left(\delta^2\right) - \theta^2 = \tau^2$. In short, $D_1$ and $D_2$ have the bivariate normal marginal distribution

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} \theta \\ \theta \end{bmatrix}, \begin{bmatrix} \tau^2 + s_1^2 & \tau^2 \\ \tau^2 & \tau^2 + s_2^2 \end{bmatrix}\right),$$

from which the following conditional distribution is obtained

$$D_2 \mid D_1 = x \;\sim \mathcal{N}(\frac{s_1^2}{\tau^2 + s_1^2}\theta + \frac{\tau^2}{\tau^2 + s_1^2}x, \frac{\tau^2 s_1^2}{\tau^2 + s_1^2} + s_2^2).$$

Let $(l_1, u_1)$ and $u_2$ be the interim and final boundaries, which means that the null hypothesis $H_0$ will be rejected (and superiority claimed) if the standardised difference observed at the interim stage is greater than $u_1$ or if the standardised final observed difference is greater than $u_2$, whereas the trial will be stopped if the standardised difference observed at the interim stage is smaller than $l_1$. Formally, $H_0$ will be rejected if either of the two disjoint events

$$\{D_1 > \delta_0 + u_1 s_1\} \text{ or } \{\delta_0 + l_1 s_1 \leq D_1 \leq \delta_0 + u_1 s_1 \cap D > \delta_0 + u_2 s\},$$

occurs. The specific values of the boundaries $l_1, u_1$ and $u_2$ may be specified, for example, by Pocock or O'Brien-Fleming rules (see for example [19]). PPS thus becomes

$$\begin{aligned}
\text{PPS} &= \mathrm{P}(D_1 > \delta_0 + u_1 s_1) + \mathrm{P}(\delta_0 + l_1 s_1 \leq D_1 \leq \delta_0 + u_1 s_1 \cap D > \delta_0 + u_2 s) \\
&= (1 - F_{D_1}(\delta_0 + u_1 s_1 | \theta, \tau^2 + s_1^2)) + \\
&\qquad \int_{\delta_0 + l_1 s_1}^{\delta_0 + u_1 s_1} \mathrm{P}(D > \delta_0 + u_2 s | D_1 = x)\; dF_{D_1}(x|\theta, \tau^2 + s_1^2) \\
&= (1 - F_{D_1}(\delta_0 + u_1 s_1 | \theta, \tau^2 + s_1^2)) + \\
&\qquad \int_{\delta_0 + l_1 s_1}^{\delta_0 + u_1 s_1} \mathrm{P}(D_2 > \frac{n}{n_2}(\delta_0 + u_2 s) - \frac{n_1}{n_2} D_1 | D_1 = x)\; dF_{D_1}(x|\theta, \tau^2 + s_1^2) \\
&= (1 - F_{D_1}(\delta_0 + u_1 s_1 | \theta, \tau^2 + s_1^2)) + \\
&\quad \int_{\delta_0 + l_1 s_1}^{\delta_0 + u_1 s_1} \left(1 - F_{D_2}(\frac{n}{n_2}(\delta_0 + u_2 s) - \frac{n_1}{n_2} x | \frac{s_1^2}{\tau^2 + s_1^2}\theta + \frac{\tau^2}{\tau^2 + s_1^2} x, \frac{\tau^2 s_1^2}{\tau^2 + s_1^2} + s_2^2)\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad dF_{D_1}(x|\theta, \tau^2 + s_1^2).
\end{aligned}$$

PPS calculations are less explicit in this case, as they are in the great majority of designs slightly more complicated than the vanilla case of "normal case with no interim analysis". In general cases, PPS is in general calculated approximately by simulation, as recommended in Spiegelhalter *et al.* (1984, page 201), or by numerical integration.

### 3.2 Interim evaluation stage

Now suppose the trial has reached the interim stage and $D_1$ has been observed to lie within the non-stopping region $\{\delta_0 + l_1 s_1 \leq D_1 \leq \delta_0 + u_1 s_1\}$. The value of $D_1$ is to be considered constant, since it has been observed. At this *interim evaluation* stage, PPS is therefore

$$\begin{aligned}
\text{PPS} &= \mathrm{P}(D > \delta_0 + u_2 s | D_1) \\
&= 1 - F_{D_2}(\frac{n}{n_2}(\delta_0 + u_2 s) - \frac{n_1}{n_2} D_1 \big| \frac{s_1^2}{s_1^2 + \tau^2}\theta + \frac{\tau^2}{s_1^2 + \tau^2} D_1, \frac{\tau^2 s_1^2}{s_1^2 + \tau^2} + s_2^2). \qquad (4)
\end{aligned}$$

Notice that the posterior on $\delta$ is still in the same normal class as the prior, due to the property of conjugacy,

The special case of a diffuse prior is of particular interest. Consider a sequence of normal priors with mean $\theta$ and such that $\tau^2 \to \infty$. In the limit, we obtain correspondingly

$$\text{PPS} = 1 - F_{D_2}(\frac{n}{n_2}(\delta_0 + u_2 s) - \frac{n_1}{n_2} D_1 \big| D_1, s_2^2),$$

an expression which does not depend on the prior. This form of noninformative PPS is very important in practice, since it gives the sponsor an "objective" evaluation of where the trial is heading. Expression (4) is to be used not to take any formal decision, but only to guide the behavior of the sponsor in the conduction of operations surrounding the trial itself. As pointed out by Spiegelhalter *et al.* (1984, page 213) "we follow Armitage (1991) in warning against using this predictive procedure as any kind of formal stopping rule. It gives an undue weight to "significance", and makes strong assumptions about the direct compatibility of future data with those data already observed".

### 3.3 Third example: a superiority clinical trial with interim analysis

Reconsider the first example and suppose now that the sponsor plans for one interim analysis, to possibly terminate the trial for early demonstration of efficacy. For example, the drug may be a new chemical agent addressing a truly urgent need for patients and the RAs are willing to grant an accelerated path to registration. At the interim stage the clinical trial may therefore be terminated when it shows convincing positive evidence; but even if it does not, it is important for the sponsor to have a prediction of the chances of final success, since parallel business plans related to the novel aspects of the drug may be affected. When planning when and how the interim analysis should be conducted, different scenarios can be compared with each other.

Assume, for example, that an interim look at 2/3 of information is plannned, i.e. when 361 patients per arm have been observed. An O'Brien-Fleming boundary rule is considered, giving interim boundaries (via tables in [19], for example, or dedicated software) $u_1 = 2.1351$ and a final boundary $u_2 = 1.6941$ ($l_1 = -\infty$ because early stopping is only allowed for demonstrated efficacy).

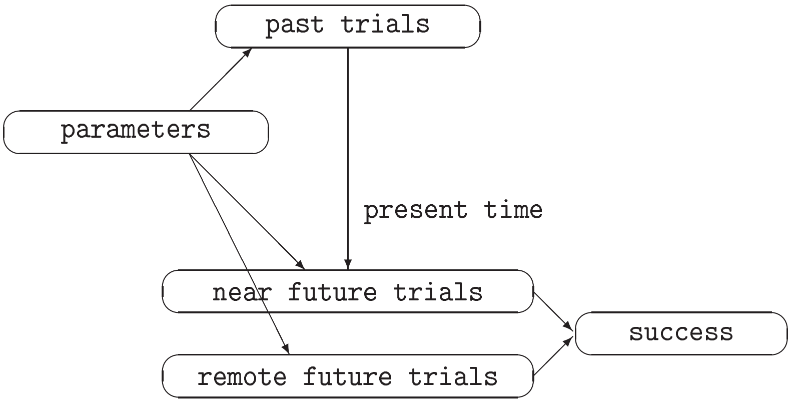
First of all, notice that PPS at planning stage is again 0.77, the same as it was in the first example, because the O'Brien and Fleming boundaries are chosen in such a way to preserve the probability of type I error.

Next, consider what may happen at evaluation stage. Suppose the observed difference at interim is $D_1 = 3$ minutes. PPS (equation (4)) equals 0.23, indicating a low chance of success.

On the non-Bayesian side, there does not exist a universally agreed quantitative measure for monitoring a clinical trial at interim, but several authors (for example [22]) give recommendations according to a "stochastic curtailment" concept. They propose as monitoring tool the *interim power* at $\delta_A = 10$, defined as the probability that, having observed $D_1 = 3$ at interim, the final outcome of the trial will be a success, given the alternative value $\delta_A = 10$. Formally, this is equivalent to PPS corresponding to a degenerate prior, which assigns probability 1 to the alternative value $\delta_A = 10$. With the same computational tools it can therefore be calculated that interim power equals $\pi(\delta_A|D_1 = 3) = 0.54$, much greater than 0.23. The larger interim power value may lead to an optimistic expectation about trial success, for the same reasons discussed for power at the planning stage. The discrepancy between the two power values arises from the fact that the interim power approach assumes the remaining 1/3 of information as coming from $\delta_A = 10$. PPS, instead, reflects our intuitive expectation of a low chance of success after having observed a mean difference of only 3 minutes at 2/3 of the way.

A second scenario the sponsor might consider is an earlier interim analysis, say at 1/3 information, i.e. 180 patients per arm at interim look. The O'Brien-Fleming boundary rule gives $u_1 = 3.2$ and $u_2 = 1.6471$. If the observed difference were $D_1 = 3$ minutes, the interim power at $\delta_A = 10$ would then be $\pi(\delta_A|D_1 = 3) = 0.86$. PPS would instead equal 0.50, thus reflecting a high level of uncertainty induced by the somewhat low observed difference. Again, the larger interim power may lead to unjustified optimism.

**Figure 1.** Graphical representation of across-trial predictions



## 4 ACROSS-TRIAL SUCCESS PREDICTIONS

Across trial PPS refers mainly to the prediction of success made at crucial points in clinical development, for example after some critical recommendations by RAs or at the go/nogo decision point between phase II and phase III. Good examples of the latter situation are [3] and especially [4], which contains a detailed account of the simulated part of a clinical development process.

A trial is usually part of a larger development program involving all kinds of scientific and economical plans. It is vital for a company to have the most up-to-date predictions on what are the chances of success of an ongoing project and to prepare as soon as possible for corrective actions, aimed at preventing risks, along the way. For example, building a new production plant is a huge economical effort and has to be planned well in advance. Other business plans that can be affected by the results of ongoing experimentation are the recruitment of additional centers for planned multi-center trials or satellite marketing studies to position a new product in the right market segment.

From a conceptual point of view, the Bayesian mechanism of updating knowledge about crucial parameters by processing all available information is applicable in the across-trial situation as well. The same parameters of interest may characterize both previous and future data, so it is legitimate for a sponsor to try to quantify the probability of success through predictive probability calculations.

Figure 1 a very general description of the approach in a pseudo-Bayesian graphical model. Past trial data and future trial data are characterized by the same parameters of interest. Future success is a (deterministic) function of future data, which may be separated into near and remote ones. At the “present” time point, a prediction is need about how future trial will result in a success or failure. The prediction will result in a PPS which will condition on past data and try to find the (posterior marginal) probability of success.

All within-trial PPS calculations shown in the previous sections can be used as ingredients to compute across-trial PPS calculations.

In practice, when trying to make predictions across trials which may be performed in very different conditions and may have different sizes and populations, it may be wise not to rely excessively on modeling and to apply some precautions, such as discounting for inter-trial uncertainty and such. Moreover, predictions which are made at very distant points in time, such as predictions involving remote future trials, should be taken with a grain of salt, keeping in mind that in any case predictions about a remote future will be updated in the light of new coming data, in a continuously improving virtuous Bayesian cycle.

### 4.1 Two large confirmatory trials after phase II

Consider, for example, a situation in which past data consists of the phase II data and future success is equivalent to the successful completions of two confirmatory trials. To follow up with the standard normal homoschedastic situations dealt with in the previous sections, suppose the results from the two large trials (either parallel or cross over, superiority or noninferiority) can be conveniently summarized by two normal sufficient statistics

$$\begin{aligned} D_1 \mid \delta &\sim \mathcal{N}(\delta, s_1^2 = c_1\sigma^2) \\ D_2 \mid \delta &\sim \mathcal{N}(\delta, s_2^2 = c_2\sigma^2) \end{aligned}$$

where $D_1$ is the result of the next trial and $D_2$ is the result of the following one. Suppose further their distributions have mean equal to an unknown treatment effect $\delta$ and known, but possibly different, variances $c_1\sigma^2$ and $c_2\sigma^2$ for some constants $c_1$ and $c_2$,. A reasonable assumption is that $D_1$ and $D_2$ are conditionally independent given $\delta$. As in the previous sections, suppose further that $\delta \sim \mathcal{N}(\theta, \tau^2)$. Following the same computations to Section 3.1, it can be verified that

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} \theta \\ \theta \end{bmatrix}, \begin{bmatrix} \tau^2 + s_1^2 & \tau^2 \\ \tau^2 & \tau^2 + s_2^2 \end{bmatrix}\right),$$

where now $D_1$ and $D_2$ refer not to two parts of the same trial but to two different trials. Suppose success is equivalent to

$$\{D_1 > \delta_{0,1} + z_{\alpha_1} s_1\} \text{ and } \{D_2 > \delta_{0,2} + z_{\alpha_2} s_2\}.$$

Then, we can compute PPS either by methods as in Section 3.1 or, equivalently, by marginalizing on $\delta$:

$$\text{PPS} = \int (1 - F_{D_1}(\delta_{0,1} + z_{\alpha_1} s_1 | d, s_1^2))(1 - F_{D_2}(\delta_{0,2} + z_{\alpha_2} s_2 | d, s_2^2)) dF_\delta(d|\theta, \tau^2),$$

which can be done either by numerical integration or by simulation.

### 4.2 Fourth example: from phase II to phase III

An example of the methods illustrated in Section 4.1 may be the transition to phase III based on a definite dose finding trial run at the end of phase II. Dose finding trials often mark the end of the explorative early drug development phases (I and II) and the totality of information accumulated until that decision point has to be carefully evaluated before running the large and costly pivotal phase III program.

Consider the phase II dose finding trial described in [23], with a total of 100 patients allocated equally to either placebo or one of four active dose levels. The response variable was assumed to be normally distributed and larger values indicated a better outcome. Assume that this outcome variable is to be used for a confirmatory phase III trial. One outcome of the phase II dose finding trial is the selection of a designated dose level to be continued (and confirmed) in phase III. The selection itself can rely on statistical reasonings (such as the MCP-Mod methodology described in [23], or any other reasonable dose finding analysis method) and include non-statistical considerations (marketing perspectives, regulatory requirements, etc. ).

The results from phase II can be the basis for constructing a prior relevant for phase III in the following way. Assume dose level 0.2 is selected for the phase III program. From Table 2 in [23], a good value for the prior mean $\theta$ of the treatment effect $\delta$ can be taken as $\theta = 0.46$. Choices of the variances are most sensitive. Since from [23] we have $MSE = 0.5$, we can set $\sigma^2 = MSE = 0.5$, while discounting the information from phase II to find a conservative value for $\tau^2$. To do so, one possibility is to take $\tau^2 = 2MSE/10$, which would be very roughly equivalent to considering the information for dose 0.2 to be provided by only 10 observations instead of 20 and ignoring that parallel groups have been run. This way, some discounting is applied due to changing conditions from phase II to phase III, but on the other hand information coming from all four experimental doses, of which only one is selected for phase III, is exploited.

## 5 CONCLUSIONS

This paper presents some applications of Bayesian predictive calculations to the context of clinical trials. Focus is on the PPS concept and calculations with normal data with known variance are spelled out in detail.

In case the sampling variance is not known, the same formulas can be used together with simulation methods, whereby many values for the unknwon variance are simulated from a prior, then PPS is computed conditional on that variance and averaged over the simulated values. This method is more efficient and realistic than using conjugate priors, which would result in Student's $t$-like calculations. The resulting Bayesian clinical trial simulation exercise has been described in detail in [1]. Here, for the sake of simplicity, we have not elaborated further on the subject.

The same reference [1] deals with some of the same material as the present paper, but the focus is mainly on single trials without interim analysis. In this paper we have rather focused on a thorough use of the simple normal case with known variance to illustrate the logic of the use of predictive probability within trials, at interim analysis and across trials. We therefore think of this paper as a tutorial-style account of some principles which have guided some of our experiences in pharmaceutical development, rather than a detailed account of case studies which would not be possible to disseminate for confidentiality reasons.

The use Bayesian tools is getting more and more common even in the highly regulated world of pharmaceutical development, as witnessed by the recent guideline [24]. Predictive probability is an important Bayesian tool from the scientific and from the business point of view. It was invented in the '80s but its literature is rich and growing, as shown by the list of references we have tried to comment.

## References

[1] A. O'Hagan, J.W. Stevens, and Campbell M.J. Assurance in clinical trial design. Pharmaceutical Statistics, 4:187–201, 2005.

[2] D.J. Spiegelhalter, K.R. Abrams, and J.P. Myles. Bayesian Approaches to Clinical Trials and Health-care Evaluation. John Wiley and Sons, Ltd, 2004.

[3] N. Stallard, J. Whitehead, and S. Cleall. Decision-making in a phase II clinical trial: a new approach combining bayesian and frequentist concepts. Pharmaceutical Statistics, 4:119–128, 2005.

[4] R. M. Nixon, A. O'Hagan, J. Oakley, J. Madan, J. W. Stevens, N. Bansback, and A. Brennan. The rheumatoid arthritis drug development model: a case study in bayesian clinical trial simulation. Pharmaceutical Statistics, 8:371–389, 2009.

[5] D.J. Spiegelhalter and L.S. Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. Statistics in Medicine, 5:1–13, 1986.

[6] D.J. Spiegelhalter, L.S. Freedman, and P.R. Blackburn. Monitoring clinical trials: Conditional or predictive power? Controlled Clinical Trials, 7:8–17, 1986.

[7] D. Johns and J.S. Andersen. Use of predictive probabilities in phase II and phase III clinical trials. Journal of Biopharmaceutical Statistics, 9:67–79, 1999.

[8] S.M. Berry, B.P. Carlin, J.J. Lee, and P. Muller. Bayesian Adaptive Methods for Clinical Trials. Chapman and Hall / CRC Press, 2010.

[9] S. Gubbiotti and F. De Santis. Classical and bayesian power functions: their use in clinical trials. Biomedical Statistics and Clinical Epidemiology, 2(3):201–211, 2008.

[10] S. C. Choi and P.A. Pepple. Monitoring clinical trials based on predictive probability of significance (c/r: V46 p274-275). Biometrics, 45:317–323, 1989.

[11] A.P. Grieve. Predictive probability in clinical trials. Biometrics, 47:323–330, 1991.

[12] B. Lecoutre. Bayesian predictive procedures for designing and monitoring experiments. In Bayesian Methods with Applications to Science, Policy and Official Statistics, pages 301–310. Luxembourg: Office for Official Publications of the European Communities, 2001.

[13] N. Dallow and P. Fina. The perils with the misuse of predictive power. Pharmaceutical Statistics, 10, 2011.

[14] D.J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6):657–680, 1987.

[15] M. D. Perlman and L. Wu. The emperor's new tests (Pkg: 355-381). Statistical Science, 14(4):355–369, 1999.

[16] Stephen Senn. Cross-over trials in drug development: Theory and practice. Journal of Statistical Planning and Inference, 96(1):29–40, 2001.

[17] J.E. Grizzle. The two-period change-over design and its use in clinical trials. Biometrics, 21:467–480, 1965. Corrigenda 30, 727, (1974).

[18] D. Spiegelhalter, L. Freedman, and M. Parmar. Bayesian approaches to randomized trials. Journal of the Royal Statistical Society. Series A, 157:357–416, 1994.

[19] C. Jennison and B.W. Turnbull. Group Sequential Methods with Applications to Clinical Trials. CRC Press, 2000.

[20] D.A. Berry. Interim analysis in clinical trials. The American Statistician, 41:117–122, 1987.

[21] A. Dmitrienko and M-D Wang. Bayesian predictive approach to interim monitoring in clinical trials. Statistics in Medicine, 25(13):2178–2195, 2006.

[22] M. Halperin, K.K.G. Lan, J.H. Ware, N.J. Johnson, and D.L. DeMets. An aid to data monitoring in long-term clinical trials. Controlled Clinical Trials, 3:311–323, 1986.

[23] F. Bretz, J.C. Pinheiro, and M. Branson. Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics, 61(13):2178–2195, 2005.

[24] Food and Drug Administration Center for Devices and Radiological Health. Guidance for the use of bayesian statistics in medical device clinical trials. Biomedical Statistics and Clinical Epidemiology, February 5, 2010.