Received: 29 June 2020 | Revised: 9 April 2021 | Accepted: 3 May 2021

DOI: 10.1002/pst.2132

MAIN PAPER

WILEY

Selection bias, investment decisions and treatment effect distributions

Stig Johan Wiklund[1] | **Carl-Fredrik Burman**[2]

[1]Captario, Gothenburg, Sweden

[2]Data Science & AI, Biopharmaceutical R&D, AstraZeneca, Gothenburg, Sweden

Correspondence
Stig Johan Wiklund, Captario, Gothenburg, Sweden.
Email: stig-johan.wiklund@captario.com

Abstract

When making decisions regarding the investment and design for a Phase 3 programme in the development of a new drug, the results from preceding Phase 2 trials are an important source of information. However, only projects in which the Phase 2 results show promising treatment effects will typically be considered for a Phase 3 investment decision. This implies that, for those projects where Phase 3 is pursued, the underlying Phase 2 estimates are subject to selection bias. We will in this article investigate the nature of this selection bias based on a selection of distributions for the treatment effect. We illustrate some properties of Bayesian estimates, providing shrinkage of the Phase 2 estimate to counteract the selection bias. We further give some empirical guidance regarding the choice of prior distribution and comment on the consequences for decision-making in investment and planning for Phase 3 programmes.

KEYWORDS

clinical development, COVID-19, efficacy prior, Go/No Go decision, regression to the mean

1 | INTRODUCTION

The article entitled "Why Most Published Research Findings Are False"[1] has been one of the most cited in recent years. Selection bias is one of the key drivers behind the provocative title. As experiments are finite in size, the estimates of key parameters will be variable. If estimates are large, they tend to be highlighted; if estimates are low, they tend to be hidden away and forgotten. There are several mechanisms contributing to this pattern: Investigators may not adjust for multiple comparisons within a trial. They are more likely to submit results that are positive, and journals have been more inclined to publish findings that appear to be important. To counteract these problems, there has lately been a strong movement to increase the transparency of clinical trial results. Regulation has been strengthened and many pharmaceutical companies have gone beyond regulations to commit to even higher degrees of transparency.

However, transparency will not completely solve the issue with selection bias. Even if corrections are made for multiple analyses within each trial, there is no obvious correction for viewing results across multiple trials and multiple drugs. A large pool of candidate drugs (CD) is tested in Phase 2, in trials with limited sample size, and the results are subject to random uncertainty. As a considerable fraction of these CDs are terminated due to lack of efficacy, it is reasonable to believe that it is common for CDs to have a lower efficacy than is anticipated in the planning phase. Among CDs selected for further development based on promising Phase 2 results, the estimated efficacy in Phase 2 is therefore

on average higher than the true efficacy. This gives rise to the selection bias that this article focuses on. Selection bias is less of a problem in regulatory decisions. The regulatory system, where promising results in Phase 2 normally have to be confirmed in Phase 3, significantly decreases the risk that non-effective drugs will reach the patients. The large investments needed in Phase 3 further limits the number of drugs that receive positive investment decisions. In this article, we will study selection bias and discuss implications for the sponsors' Phase 3 investment decisions.

Several authors have studied selection bias (e.g. Efron[2]) and identified that this bias is an important factor behind that positive results in Phase 2 are often not replicated in Phase 3.[3-5] This can be viewed as a regression to the mean mechanism and will be driven by the distribution of efficacies among the wider pool of CDs tested. For the specific case that the estimate from the Phase 2 trial is used to plan the upcoming Phase 3 trial, several methods to adjust the Phase 2 estimates have been proposed.[6,7] Chuang-Stein and Kirby[5] highlight three reasons why this is important, in that it impacts our ability to make appropriate decisions:

- Going into Phase 3 development based on overly optimistic estimates can lead to underpowered, and eventually failed, confirmatory programmes.
- Endpoint selection can be led astray if one endpoint is selected for the confirmatory programme on the basis of a randomly high estimate.
- There is a risk that a suboptimal subgroup is selected as the primary analysis population in Phase 3.

The fact that Phase 2 and Phase 3 trials may often provide divergent results has also been addressed by the FDA.[8] In a case study of 22 projects, they conclude that "phase 2 results can inaccurately predict safety and/or effectiveness for medical products in a wide range of diseases and patient populations." In a recent article, Qu et al.[9] present an approach to the adjustment for selection bias, which is illustrated for the situation of a transition between an early PoC study and a larger Phase 2 study, and the approach is applied to some projects from the diabetes and immunology disease areas. While our work is instead focussed on the transition between Phase 2 and 3, it has similarities to the situation studied by Qu et al. as in both cases the results from a smaller early study may be subject to selection bias when proceeding to a subsequent larger study.

To evaluate the degree of the selection bias occurring at Phase 3 investment decisions, and to study potential consequences, we argue that it is essential to apply a relevant distribution for treatment efficacies of CDs to account for the uncertainty pertaining to the efficacy of the drug going into Phase 2. While many authors have performed this type of evaluation considering scenarios where the treatment effect has been set to a fixed value,[6,10] we choose to extend the approach taken by Chuang-Stein and Kirby[11] in applying a prior distribution to the treatment effect.

Chapter 2 suggests three examples of different types of prior distributions: a mixture distribution with significant probability for no efficacy, a lognormal distribution and an (improper) exponential distribution. In Chapter 3, we consider the sub-probability distributions of CD efficacies, resulting from a selection mechanism used for the Phase 3 investment decision. In addition, the conditional distributions and means of Phase 2 estimates, given a Phase 3 go decision, are provided. Utilising an assumed prior distribution, Chapter 4 then considers discounting of Phase 2 results, producing a posterior and an assessment of the selection bias. We also study robustness properties when the prior distribution is misspecified. The three different priors introduced in Chapter 2 are used throughout Chapters 3 and 4. Ideally, the choice of prior should be informed by empirical data. In Chapter 5, we therefore extract data from clinicaltrials.org in order to illuminate the choice of prior distribution. An illustrating example is provided in Chapter 6, where we exemplify the issue of selection bias with the case of testing medicines for the COVID-19 disease. Results and methods are discussed in Chapter 7.

2 | TREATMENT EFFECT DISTRIBUTIONS

A large number of CD are tested in Phase 2, by different sponsors. For drugs with similar endpoints and for similar indications, we can think of an underlying distribution of efficacy. We assume that the efficacy of an arbitrary drug going into Phase 2 follows a distribution π. It may be possible to estimate this distribution using historic data on observed Phase 2 efficacy for a large number of drugs. Alternatively, we can view π as reflecting the beliefs and uncertainty for a particular drug, in which case π may be interpreted as a prior distribution in a Bayesian sense.

Assume that the efficacy of an arbitrary drug going into Phase 2 follows a distribution $\pi(\theta)$, where θ is the true efficacy of the drug. Three different classes of distributions for $\pi(\theta)$ will be considered and used to illuminate the issues at

hand. These classes have been chosen to highlight various types of behaviours, for example also the possibility of a *negative* selection bias, and to study the robustness of conclusions with respect to different prior assumptions.

1. Mixture of a point mass at zero and a normal distribution

$$\pi(\theta) = \begin{cases} q\, d(0) & \text{if } \theta = 0 \\ (1-q)\left(2\pi\tau_N^2\right)^{-1/2} exp\left(-\frac{(\theta-\mu_N)^2}{2\tau_N^2}\right) & \text{if } \theta \neq 0 \end{cases}$$

where $d(0)$ is the Dirac function. That is, the probability of efficacy being exactly zero is assumed to be q. Given non-zero efficacy, a normal distribution is assumed. This is an approach taken for instance by Chuang-Stein and Kirby [11, Ch 5]. The rationale being that the lump probability at zero would represent the proportion of drugs historically shown to have no (or very small) effect, and the proportion assigned to the normal distribution would be centred around a historical average for efficacious drugs.

2. Exponential distribution

$$\pi(\theta) = k * exp(-k\theta)$$

An exponentially distributed prior was for instance used by Miller and Burman.[12] A rationale for this prior is that it might be reasonable to assume that low or moderate efficacies are more likely than high ones. A non-ignorable number of CD are likely to have only marginal efficacy. This indicates that the prior density for efficacy should perhaps be a decreasing function.

3. Log-normal distribution

$$\pi(\theta) = \left(2\pi\theta^2\tau_L^2\right)^{-1/2} exp\left(-\frac{(\ln\theta-\mu_L)^2}{2\tau_L^2}\right)$$

A rationale for a log-normal prior could be that the process through discovery and pre-clinical development should have successfully terminated drug candidates which are totally void of mechanistic action towards efficacy. The candidates progressed to clinical development might still have substantial likelihood of sub-optimal efficacy, but the likelihood close to zero should be negligible. The log-normal distribution was proposed as a model for the true treatment effect by Wiklund.[13]

For simplicity, the distributions are selected to yield the same mean ($\mu_\pi = 1$), using the following parameter values: $p = 0.5$, $\mu_N = 2$, $\tau_N = 0.6$, $k = 1$, $\mu_L = -0.125$, $\tau_L = 0.5$.

3 | DISTRIBUTION AFTER SELECTION FROM PHASE 3 INVESTMENT DECISION

At the investment decisions made after Phase 2, the sponsor typically terminates those projects that are deemed not to exhibit a sufficient treatment effect, and only progress to Phase 3 those projects showing good efficacy. This leads to selection bias; conditioning on Phase 3 progression, the observed treatment effect in Phase 2 is a biased prediction of the Phase 3 results (e.g., Qu et al.[9]). In this chapter, the impact of selection at an investment decision will be illustrated based on the three prior distributions introduced above.

Assume that the efficacy estimate from Phase 2 is based on the comparison of the mean of two treatment arms, for example active treatment versus placebo. If we further assume that the estimate is normally distributed, we have that $\hat{\theta}_2 \,|\, \theta \sim N(\theta, \tau^2)$, where $\tau = \sigma\sqrt{2/N}$ is the standard error of the Phase 2 estimate and N is the sample size per treatment arm. Following the Phase 2 trial, a stop/go decision is made whether to terminate the project or to progress to Phase 3. There are different ways of specifying how large the observed Phase 2 efficacy has to be to invest in Phase 3 (see e.g., Frewer[14]). In this article, we have chosen a stop/go criterion based on whether efficacy is statistically significant in Phase 2. Denote the z-score by $Z = \hat{\theta}_2/\tau$ and write C for the critical value. The event of Phase 3 transition is then $Q = \{Z > C\}$. The probability for this is the statistical power.

$$P_N\left(Q|\theta\right)=\Phi(Z-C)=\Phi\left(\frac{\theta}{\sigma}\sqrt{\frac{N}{2}}-C\right)$$

The sub-probability density of efficacy for the subset of projects progressed to Phase 3 is obtained by multiplying the prior distribution (as defined in Ch 2) with the statistical power, $p_N^*(\theta)=\pi(\theta)\cdot P_N(Q|\theta)$. Scaling this by the probability of Phase 3 progression, $P_N(Q) = \int \pi(\theta)\cdot P_N(Q|\,\theta)d\theta$ we get the efficacy density given progression, $p_N\left(\theta|Q\right)=p_N^*(\theta)/P_N(Q)$. In Figure 1, $p_N^*(\theta)$ for each of the three priors are shown for different sample sizes, $N = 20$, 50, 100, 200.

The graphs in Figure 1 illustrate the extent to which the selection mechanism of the investment decision implies a different distribution after progression to Phase 3, as compared to the prior. (Note that the *y*-axis varies between graphs.) The selection does to some extent succeed in selecting the projects with higher efficacy, hence shifting the distribution of successful projects to the right. This selection is however by no means perfect. As illustrated in the graphs, there will be a substantial probability that progressed projects will have a true efficacy that is probably lower than anticipated (low values of θ are part of the selected distribution). On the other hand, many projects with highly efficacious drugs will be terminated (the selected distribution is substantially lower than the prior even for relatively high values of θ). The degree to which the distribution is shifted does obviously depend on the sample size, *N*. Larger sample sizes will lead to more accurate decisions, that is less false negative and false positive investment decisions. The impact of the sample size is however different between the priors. The shape of the selected distributions from the mixture prior appears to be relatively similar, independent of *N*. The selected distributions for the other priors depend more clearly on sample size.

We may calculate the mean of the treatment effect for the subset of projects progressed to Phase 3,

$$E_N\left[\theta|Q\right]=\int\theta\cdot p_N(\theta|Q)d\theta.$$

Given that a drug with efficacy θ progresses to Phase 3, the density of the Phase 2 estimate follows a normed truncated normal distribution,

$$p_N\left(\hat{\theta}_2|Q,\theta\right)=\varphi\left(\hat{\theta}_2/\tau\right)\cdot 1_Q/\Phi(\theta/\tau-C),$$

where φ is the normal density and 1_Q is the indicator function for progression, that is $\hat{\theta}_2 > C\tau$. Figure 2 displays the conditional Phase 2 mean, $E_N\left[\hat{\theta}_2|Q,\theta\right]$, as an illustration of the bias conditional on transition. The results show that the conditional Phase 2 estimate severely overestimates the true treatment effect. This is particularly the case for studies with a small sample size, but the bias is also substantial for larger trials with small to moderate true treatment effects. Conditional on transition to Phase 3, we present in Table 1 some numerical results for mean efficacy, $E_N[\theta|\,Q]$, and mean Phase 2 estimate

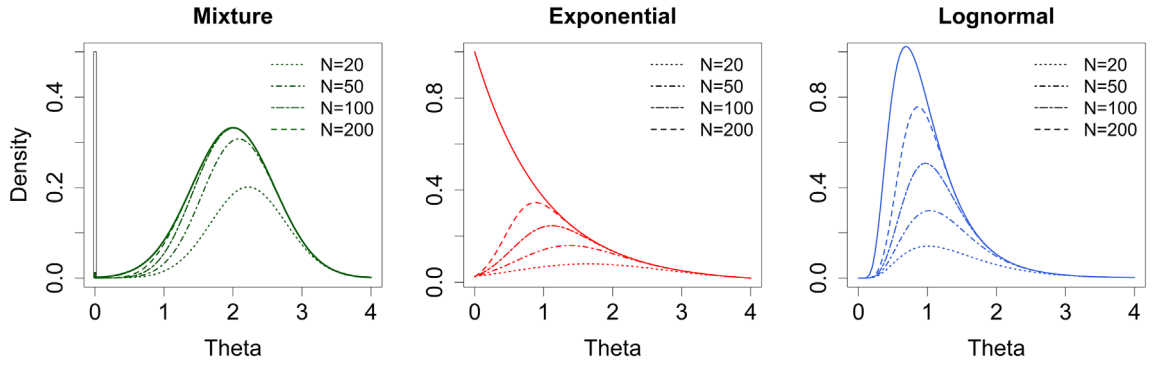

FIGURE 1 Distributions of the subset of projects with a successful Phase 3 transition, based on the three prior distributions. (The curves are representing the product of the prior and the power function and are hence improper densities with an area ≠1)

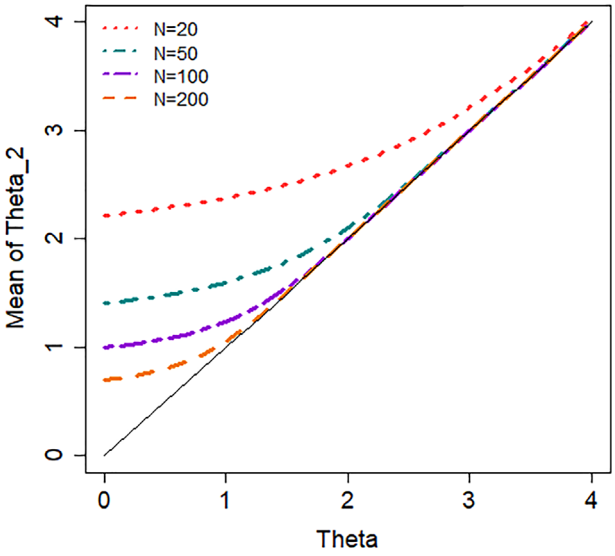

FIGURE 2 Mean of the Phase 2 estimate conditional on successful Phase 3 transition as a function of θ

TABLE 1 The average estimated efficacy, as compared to the average true efficacy, for projects with a successful Phase 3 investment decision

		Prior distribution			Value of Ph2 estimate required for Ph 3 transition
		Point mass + normal	Exponential	Log-normal	
Mean of prior distribution		**1.00**	**1.00**	**1.00**	
Mean of Ph 2 estimate, given progress to Phase 3	$N = 20$	2.79	2.82	2.52	1.86
	$N = 50$	2.32	2.14	1.79	1.18
	$N = 100$	2.06	1.80	1.44	0.83
	$N = 200$	2.00	1.56	1.22	0.59
Mean of true effect, given progress to Ph 3	$N = 20$	2.08	2.10	1.42	1.86
	$N = 50$	2.01	1.86	1.33	1.18
	$N = 100$	1.96	1.67	1.24	0.83
	$N = 200$	1.92	1.51	1.14	0.59
Probability of progress to Ph 3	$N = 20$	0.29	0.23	0.21	1.86
	$N = 50$	0.43	0.36	0.39	1.18
	$N = 100$	0.48	0.47	0.57	0.83
	$N = 200$	0.50	0.58	0.75	0.59

$$E_N\left[\hat{\theta}_2|Q\right] = \int E_N\left[\hat{\theta}_2|Q,\theta\right] \cdot p_N(\theta|Q)d\theta.$$

The results may be compared to the mean of the prior distribution, which is set to 1 for all the three priors. The results give an overall summary of the amount to which the positive investment decision selection will have an impact on the expected true treatment effect. If the treatment effect follows the mixture prior, the impact will be large, with a mean of the selected distribution being around 2 for all the evaluated sample sizes. This is due to the fact that the selection in this case mainly acts to weed out the zero part of the prior. For the other priors, the impact of the selection is smaller. For the lognormal prior, the mean treatment effect after the investment decision is relatively close to the prior mean, at least for larger sample sizes.

The results also indicate that, as expected, there is a substantial bias in the estimate from Phase 2, when conditioned on a successful Phase 3 investment decision. This bias is large for smaller sample sizes, but much smaller with a larger sample size, as a larger N reduces the variability of the estimate. As seen in Table 1 for $N = 200$, $E_N\left[\hat{\theta}_2|Q\right]$ is only slightly higher than $E_N[\theta|\, Q]$.

From results in Table 1 it may also be noted that there is a clear difference in the behaviour of the probability of study success for the three prior distributions. If the treatment effect follows the mixture prior, the success probability does not increase substantially for large sample sizes. Instead, this probability approaches a plateau given by the probability, q, of the zero part of the prior and of the type 1 error, α. For the lognormal prior, on the other hand, the success probability increases rapidly over a relevant range of Phase 2 sample sizes. It may also be noted that the success probability of the different scenarios corresponds to the area under the curves of Figure 1.

4 | BAYESIAN DISCOUNTING OF PHASE 2 ESTIMATES

4.1 | Posterior distributions

The prior distributions presented in Ch 2 represent sceptical priors in the sense that they would typically have a lower average than the efficacy anticipated in the target profile for the drug or anticipated in the traditional sample size calculations done by the sponsor. This is motivated by the fact that many drugs fail due to insufficient efficacy. The Bayesian estimate obtained from applying the prior distributions then generally represents a shrinking of the estimate towards a smaller effect estimate. The specific mixture prior may however give a higher predicted mean than the estimate, in some cases. The degree of shrinking would be given by the posterior distributions, which we will derive in this section.

4.1.1 | Mixture of a point mass at zero and a normal distribution

With the Phase 2 estimate being normally distributed, $\hat{\theta}_2 \,|\, \theta \tilde{N}(\theta, \tau^2)$, the posterior distribution for the mixture prior is

$$\pi\left(\theta|\hat{\theta}_2\right) = K \cdot \varphi\left(\frac{\hat{\theta}_2 - \theta}{\tau}\right) \cdot \left(q \cdot d(\theta) + (1-q) \cdot \varphi\left(\frac{\theta - \mu_N}{\tau_N}\right)\right)$$

where K is the norming constant. The posterior distribution when $\hat{\theta}_2 = 1$ is illustrated in Figure 3(A), for a range of sample sizes, N. As indicated in the figure, the posterior is a mixture of a normal distribution and a point mass in $\theta = 0$. When Phase 2 data are promising, the probability of no efficacy is typically decreased compared to the prior. The mean of the normal distribution part is a weighted average, reflecting amount of information, of the prior mean and the Phase 2 estimate.

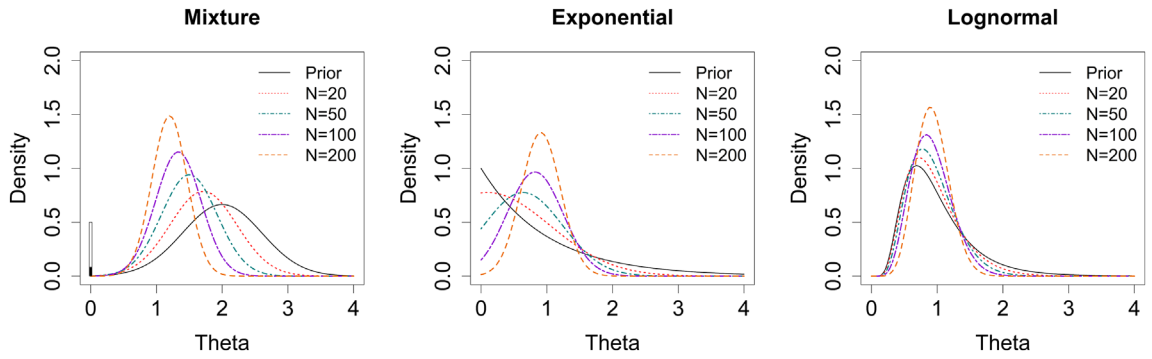

FIGURE 3 Posterior distributions for different sample sizes after updating from the three different priors. ($\hat{\theta}_2$=1. For the mixture normal prior, the height of the posterior point mass at zero represents the case of $N = 50$.)

4.1.2 | Exponential distribution

The exponential distribution is usually defined only for non-negative values, $\theta > 0$. However, the improper version of this prior, allowing any real value for θ will give a mathematically tractable posterior and provide a convenient rule-of-thumb. We will first derive the posterior and then return to discuss the improper prior. With the Phase 2 estimate, $\hat{\theta}_2$, being normally distributed, and using the improper exponential prior, the posterior is proportional to

$$\pi(\theta|\hat{\theta}_2) \propto \pi(\theta) \cdot p_N(\hat{\theta}_2|\theta) \propto exp(-k\theta) \cdot exp\left(-\frac{(\hat{\theta}_2-\theta)^2}{2\tau^2}\right) = exp\left(-\frac{(\theta-(\hat{\theta}_2-k\tau^2))^2+2\tau^2k\hat{\theta}_2-(k\tau^2)^2}{2\tau^2}\right)$$
$$\propto exp\left(-\frac{(\theta-\mu_2)^2}{2\tau^2}\right)$$

where $\mu_2 = \hat{\theta}_2 - k\tau^2$. This means that the posterior for θ is normally distributed, just like the sample distribution. Rather surprisingly, the posterior variance is identical to the sample variance, although we in some sense have added information in terms of an informative prior. If $k = 0$, the prior is so-called non-informative, and the posterior will not adjust the Phase 2 estimate. If $k > 0$, the posterior mean $\mu_2 = \hat{\theta}_2 - k\tau^2$ adjusts the estimate by an amount proportional to the sample variance. This finding may provide a convenient rule of thumb for the discounting of Phase 2 estimates.

We have in the derivation of the posterior used an improper exponential distribution, with support on $(-\infty, +\infty)$. This will lead to negative estimates when Phase 2 data are close to zero, and negative Phase 2 estimates will in fact be further amplified by the Bayesian updating. This is often unrealistic as it is less likely that the drug will have a clear negative effect compared to placebo. It might be argued that this will have little practical importance, as drugs with low or negative estimated efficacy will anyway likely not be considered for Phase 3 testing, and the rule of thumb may be useful for the majority of relevant situations. Since the mean of the (proper) prior exponential distribution is $\mu_\pi = 1/k$ and the standard error is $\tau = \sigma\sqrt{2/N}$, it follows that the adjustment applied to the Phase 2 estimate can be approximated as

$$k\tau^2 = \frac{2\sigma^2}{N\mu_\pi}$$

The posterior distribution is illustrated in Figure 3(B), for a range of sample sizes, N, and with $\hat{\theta}_2 = 1$. In the numerical illustrations of Figure 3 and in subsequent parts of the article, we only illustrate the case of $\theta > 0$ corresponding to the proper exponential prior.

4.1.3 | Log-normal distribution

With the Phase 2 estimate being normally distributed, $\hat{\theta}_2 \,|\, \theta \tilde{N}(\theta, \tau^2)$, the posterior distribution for the log-normal prior is proportional to

$$\pi(\theta) \cdot p_N\left(\hat{\theta}_2|\theta\right) \propto \frac{1}{\theta} \cdot exp\left(-\frac{(\ln\theta-\mu_L)^2}{2\tau_L^2} - \frac{(\hat{\theta}_2-\theta)^2}{2\tau^2}\right)$$

The posterior distribution is illustrated in Figure 3(C), for a range of sample sizes, N, and with $\hat{\theta}_2 = 1$.

4.2 | Properties of posterior mean estimates

From the posterior distributions, we have calculated the posterior mean,

$E_N[\theta|\hat{\theta}_2] = \int \theta \cdot \pi(\theta|\hat{\theta}_2) d\theta$. This would represent the Bayesian adjusted estimate, available for the Phase 3 investment decision, after updating with the Phase 2 results. In Figure 4 we present results showing the relationship

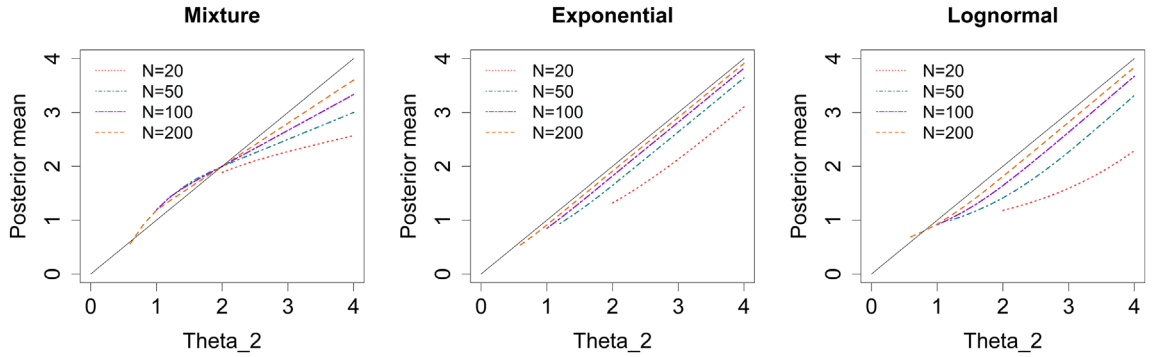

FIGURE 4 Posterior mean as a function of the Phase 2 estimate for the three different priors

between the posterior mean and the Phase 2 estimate, $\hat{\theta}_2$. The posterior mean curves are only shown for $\hat{\theta}_2 > C\tau$, i.e. for values of $\hat{\theta}_2$ large enough to lead to a positive Phase 3 transition. Figure 4 shows that the posterior mean in most cases implies a shrinkage, in that the posterior mean is lower than the Phase 2 estimate. There is however a clear difference between the prior distributions. As seen in Figure 4a, for the mixture prior, the posterior mean can in fact be higher than the Phase 2 estimate. Only for large estimates or small sample sizes, does the posterior mean imply a shrinkage.

For the exponential prior, as illustrated in Figure 4(B), the posterior mean is uniformly lower than the Phase 2 estimate. As noted in Section 4.1, the posterior mean implies a reduction from the Phase 2 estimate that is proportional to the sample variance, that is $E\left[\theta|\hat{\theta}_2\right] = \hat{\theta}_2 - k\tau^2$. This result is represented in Figure 4(B) by the curves for posterior means being approximately linear and are parallel to the Phase 2 estimate (i.e., the identity line).

The lognormal, Figure 4(C), is the prior distribution for which the posterior mean implies the largest degree of shrinkage. The amount of shrinkage is larger for smaller sample sizes, and for larger Phase 2 estimates.

4.3 | Selection bias and misspecified prior assumptions

Until now we have implicitly assumed that the prior distribution for the treatment effect is known. This implies that the prior distribution assumed in an analysis is the same as the true distribution from which the treatment effect of the drug candidate is drawn. In practical applications, the true prior distribution would be unknown. Analyses and decisions will have to be made based on an assumed prior distribution, which may be more or less incorrect.

We will in this section provide some illustrations of the consequence of this misspecification of the assumed prior. In doing so, we introduce the notation $\pi^*(\theta)$ for the assumed prior, whereas $\pi(\theta)$ denotes the true prior. Likewise, $\pi^*\left(\theta \,|\, \hat{\theta}_2\right)$ and $\pi\left(\theta \,|\, \hat{\theta}_2\right)$ are the posteriors based on the assumed and true priors. The corresponding posterior means are denoted μ_{π^*} and μ_π, respectively. The bias of the Phase 2 estimate conditional on the positive Phase 3 investment decision, that is the selection bias, is then $\beta_N\left(\hat{\theta}_2\right) = \hat{\theta}_2 - \mu_\pi$. From the graphs in Figure 4, the selection bias is represented by the difference between the identity line and the posterior mean curves. In Figure 4 we presented these results to illustrate the amount of shrinkage imposed by the posterior mean. By transforming the scale to represent bias, as given by the previous equation, the results are showing the bias of the Phase 2 estimate. We illustrate this in Figure 5, giving the selection bias, $\beta_N\left(\hat{\theta}_2\right)$, as a function of the Phase 2 estimate, for the log-normal prior and for a range of sample sizes. Figure 5 illustrates that the bias is larger for small sample sizes and for larger Phase 2 estimates.

When $\pi^*(\theta)$ is not the same as $\pi(\theta)$, the bias of the posterior mean is $\beta_N^*\left(\hat{\theta}_2\right) = \mu_{\pi^*} - \mu_\pi$, that is the difference between the posterior mean from the assumed prior and the posterior mean from the true prior. We illustrate this bias in Figure 6, taking the log-normal to be the true distribution of the treatment effect, and showing the bias of the posterior mean if incorrectly assuming the mixture normal or exponential prior, respectively. We also include in these graphs the selection bias of the Phase 2 estimate, $\beta_N^*\left(\hat{\theta}_2\right)$, noting that this may also be considered as the bias from assuming a non-informative prior in a Bayesian analysis.

The results presented in Figure 6(A) show that for a small sample size ($N = 20$) the posterior mean provides a substantial reduction in the bias of the Phase 2 estimate, even when the posterior mean is based on a misspecified prior

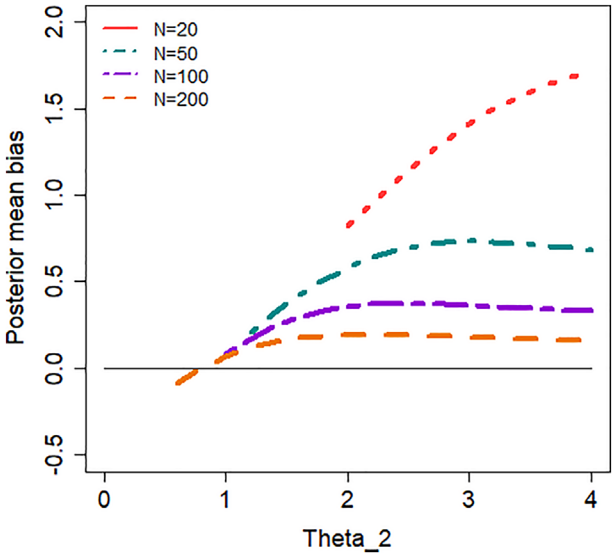

FIGURE 5 Bias in the Phase 2 estimate, when treatment effect follows a log-normal prior ($N = 20, 50, 100, 200$)

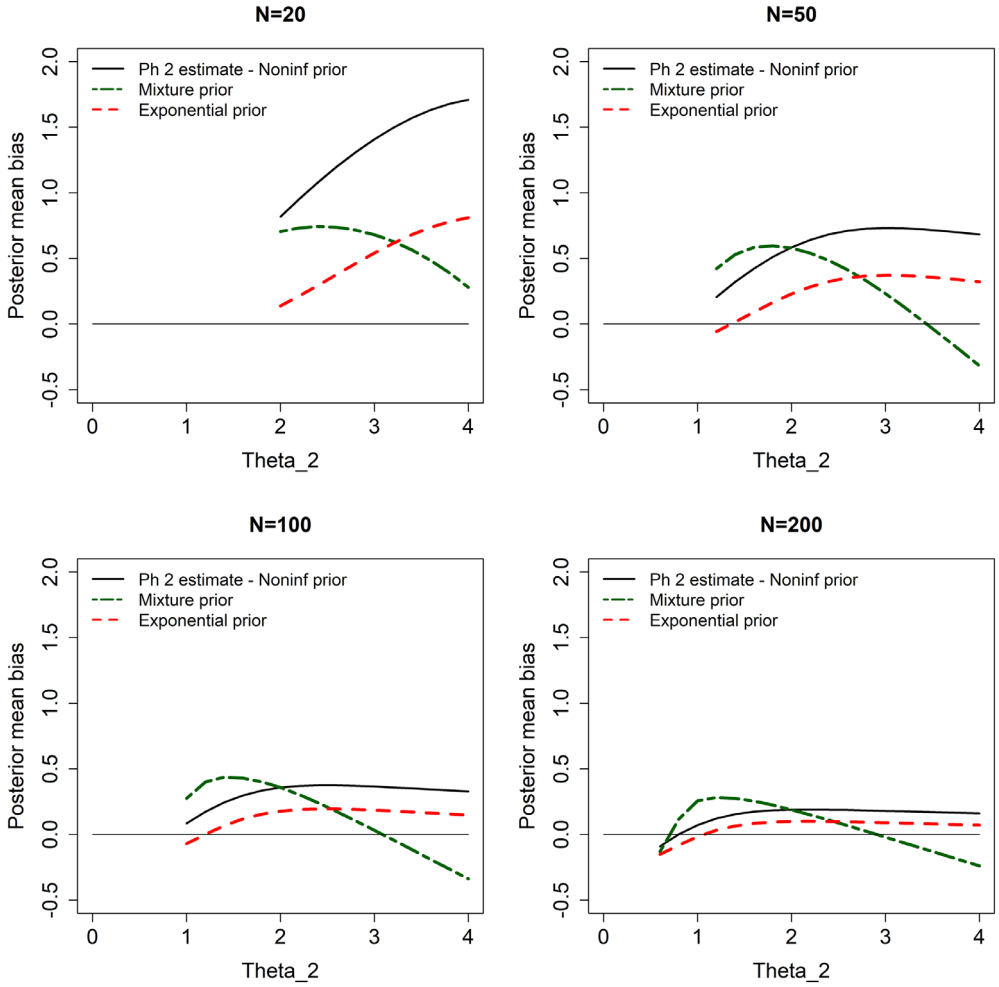

FIGURE 6 Bias in the posterior mean, as a function of the Phase 2 estimate, when the assumed prior is misspecified and the true prior is log-normal

distribution. Figures 6(B)–D) further illustrate that an analysis based on the mixture normal prior, when the true prior is lognormal, can sometimes imply a potentially unexpected behaviour. For high values of the Phase 2 estimate, the posterior mean will in fact provide an overcorrection of the selection bias and result in a negative bias for the posterior.

For low values of the Phase 2 estimate, the posterior mean will instead imply an exaggeration of the bias, by being larger than the nominal Phase 2 estimate. Figure 6 also visualises the result presented in Section 4.1, that when applying the exponential prior the posterior mean implies a constant reduction, irrespective of the value of the Phase 2 estimate. In the graphs, this is represented by the dashed curve being a parallel downward shift of the solid curve. The results indicate that using the rule of thumb based on the exponential prior does in fact imply a substantial bias correction in this case of prior misspecification.

5 | EMPIRICAL DATA

We have shown in previous sections that the impact of selection bias, and the performance of various estimates, will be highly dependent on the prior distribution of the treatment effect. To get some empirical insight into the distribution for drugs under development, we extracted data from the database available at ClinicalTrials.gov.[15] The database was searched for all industry sponsored studies in Phase 2, and the sample size and observed p-value of the primary statistical analysis were obtained for all studies where these parameters were reported. The search yielded data from $n = 1077$ trials.

For each trial, i, we calculated an approximation of the normalised effect size $\hat{\Delta}_i^*$. The calculations were in line with the assumption made earlier, that the statistical analysis in a trial can be approximated by the comparison of two means. The calculation of $\hat{\Delta}_i^*$ was done to enable a summary of the various studies and allowing their outcomes to be presented on a common scale. Details of these calculations, and of the approximations involved, are given in the Supplementary material. The distribution of the approximated effect sizes, $\hat{\Delta}_i^*$, is illustrated in the histogram of Figure 7.

We also categorised the included trials into seven different disease areas. The categorised analysis was performed since it might be argued that substantial discrepancies between disease areas would invalidate analyses on the entire data set. The distributions for each disease area are illustrated in similar histograms in the Supplementary material. It turns out that the general patterns for the effect size distributions appear to be fairly consistent across the disease areas. The lowest mean effect size is found for the Oncology and CNS disease areas, whereas the highest mean effect size is found for the disease area containing GI and metabolic disorders.

In Figure 7, we have included an approximate illustration of what distributions would have been expected based on the three different prior distributions. These were obtained by calibrating the mean of the prior distributions to equal the mean of the empirical data. A normally distributed variable was then added to each of the prior distributions, to mimic the fact that the empirical data represent both the actual distribution of true treatment effect and an additional observation error from the trials. Further details on these calculations are given in the Supplementary material, where the illustrative outcomes for the three priors are also given for each of the disease areas. We appreciate that the calculations do not represent an exact estimation procedure, but the approximations should be sufficient for the illustrative purposes of the Figures. The results indicate that different prior distributions can reasonably represent the empirical

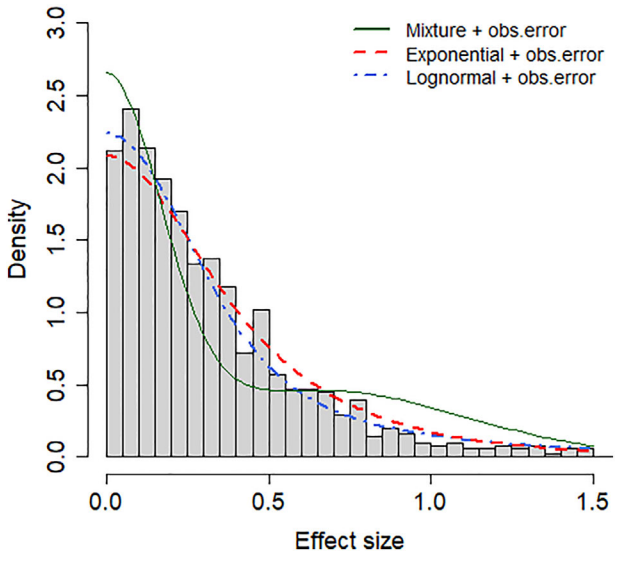

FIGURE 7 Histogram showing the distribution of the observed treatment effect size, $\hat{\Delta}_i^*$, in Phase 2, for trials with p-values and sample sizes reported in clinicaltrials.gov. Curves are overlaid to illustrate expected outcome for the three different prior distributions

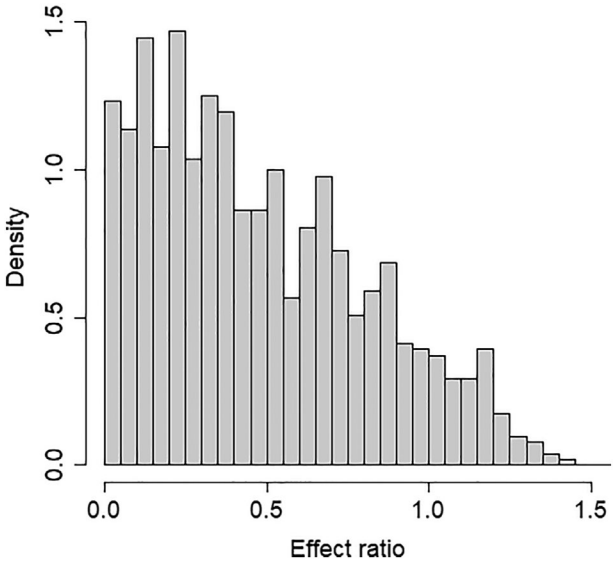

FIGURE 8 Histogram showing the distribution of the observed effect ratio, R_i^*, in Phase 2, for trials with *p*-values reported in clinicaltrials.gov

data. From our evaluated priors, the lognormal and exponential distributions seem to align relatively well to the data, whereas the mixture distribution provides a less accurate fit.

We further estimated the effect sizes, $\tilde{\Delta}_i^*$, that might have been anticipated in the planning of the trial. By calculating the ratio $R_i^* = \hat{\Delta}_i^* / \tilde{\Delta}_i^*$, we get an indication of the degree to which the observed effect size matched the effect size anticipated in the planning stage. The distribution of the observed effect ratio, R_i^*, is illustrated in Figure 8. A large part of the distribution of R_i^* is below one, which indicates that in a majority of the trials the observed treatment effect was lower than anticipated at the planning stage. It should be noted, however, that there are apparent methodological issues related to the empirical results of this section. These issues will be further addressed in the Discussion.

6 | ILLUSTRATING EXAMPLE: TESTING EXISTING DRUGS FOR COVID-19

At the time of writing this article, the SARS-CoV-2 virus pandemic is ongoing, the COVID-19 is threatening to kill millions of people and has caused severe disruptions in societies all over the globe. The acute medical need has triggered the testing in this new disease of a multitude of old pharmaceuticals, developed and sometimes approved for other indications. The medical understanding is rapidly developing and any predictive model for the outcome of these trials will soon become outdated. Still, we will use this setting as an illustration of how selection bias is generated and what the magnitude could be. In doing so, we will not make any claims of accurately modelling the current situation. Some comments on where the set-up may differ from reality are given at the end of this Chapter.

Say that 200 drugs are tested in COVID-19 versus standard-of-care (SOC). We take the true mortality on SOC to be 20% in a population of hospitalised patients. This death rate approximates early estimates (Zhou et al.[16] Richardson et al.[17]), from the time when many clinical trials were initiated. However, mortality among hospitalised patients has varied substantially depending on covariates[17] and has declined over time.[18,19] In normal drug development, CD entering Phase 2 will almost invariably have positive pre-clinical and perhaps clinical signals for the disease at hand. The already existing drugs tested for COVID-19 are less likely to be effective in this new setting, as pre-clinical data for this disease are often lacking. In this example, we are therefore assuming that 75% of tested drugs have absolutely no positive effect on mortality. Some of them could actually be harmful, although we will ignore this possibility for the current example. Thus, we take mortality = 20% for all these drugs. Among drugs that have positive efficacy (mortality strictly lower than 20%), we assume, following Collignon et al.[20] that the prior probability density is proportional to the cube of the true mortality rate. This means that it is $2^{3+1} = 16$ times more likely that an effective drug has mortality in the interval 10–20% as that its mortality is 5–10%. It should be stressed that this constitutes *one* possible subjective Bayesian prior. When considering a specific drug, it may be possible to use expert elicitation and available information (e.g., data on response biomarkers, and data on similar drugs) to tailor the prior.

With many drugs competing for research resources, early phase sample sizes are often small. We take a Phase 2 trial with $n = 50$ hospitalised patients per arm. In order to limit the Type 2 error, a mortality benefit is tested on a higher than usual significance level, alpha = 10% one-sided, using Fisher's exact test. Only drugs having a significant benefit will proceed to confirmatory Phase 3 testing. Figure 9 displays the distribution of mortality rates for drugs being effective (blue curve). It also displays the density of drugs being selected to proceed to Phase 3 (red curve) based on a statistical significance in Phase 2. This is the prior density times the power. Note that this is a sub-probability distribution, that is, the total probability that the drug gets a go decision is less than one. In this case, on average 18% of drugs with any efficacy will be selected for Phase 3. As Fisher's exact test is discrete, and sample sizes are small, the probability that a non-effective drug will move to Phase 3 is 6%, substantially smaller than the nominal alpha. However, with 75% of the tested drugs being non-effective, the expected number of non-effective drugs entering the next phase is nine. This is similar to the expected number of effective drugs. Thus, about half of drugs entering Phase 3 can be expected to be truly effective, and some of them may have insufficient efficacy to succeed in the confirmatory trial.

As an illustration, let us first condition on a drug having mortality = 15%. With the assumptions above, it is more likely than not that it does not proceed to Phase 3. Given that the drug has mortality = 15% and that it is statistically significant in Phase 2, its expected mortality in Phase 2 is only 10%. That is, the selection bias for the mortality estimate is 15–10% = 5%. Let us now remove the conditioning and consider the total population of drugs. Based on the prior and the Phase 3 go criterion, the expected selection bias among all significant drugs is 6.1% points.

The set-up we have been using differs from the COVID-19 situation in a number of ways. For example, the sample sizes in Phase 2 will obviously vary greatly between different trials and the power may depend on the choice of endpoint. As was shown in previous chapters, the sample size will have a large impact on the degree of selection bias. Some drugs may go directly into Phase 3. The two phases may also be combined into a seamless Phase 2/3 trial. The issue of selection bias will be present in similar ways as for a separate Phase 2 trial if a go/no go decision is based on the Phase 2 component of a seamless design. Many drugs are tested in platform trials, with several drug candidate sharing a common control group. This will introduce a correlation but not drastically change the issue of selection bias. A practically important aspect is that several drug candidates are related, belonging to the same class or having similar modes of action. A sophisticated predictive model could therefore use a prior with dependencies between the efficacies of such drugs. It may even be possible to borrow information between different drugs.

Although the model we have considered is not perfectly reflecting the COVID drug testing situation, it can underline the need for different stakeholders to carefully make trial decisions and to interpret results from small trials with caution. The public and journalists should ideally be taught to consider not only point estimates but also confidence intervals and p-values. As a very large number of COVID trials have been launched, a Phase 2 p-value of 0.001, say, may not be convincing. On the other hand, a p-value of 10^{-6} from a well-designed although small trial of a drug with clinically plausible efficacy can be interpreted as proof of efficacy, although the point estimate (and CI) should not be

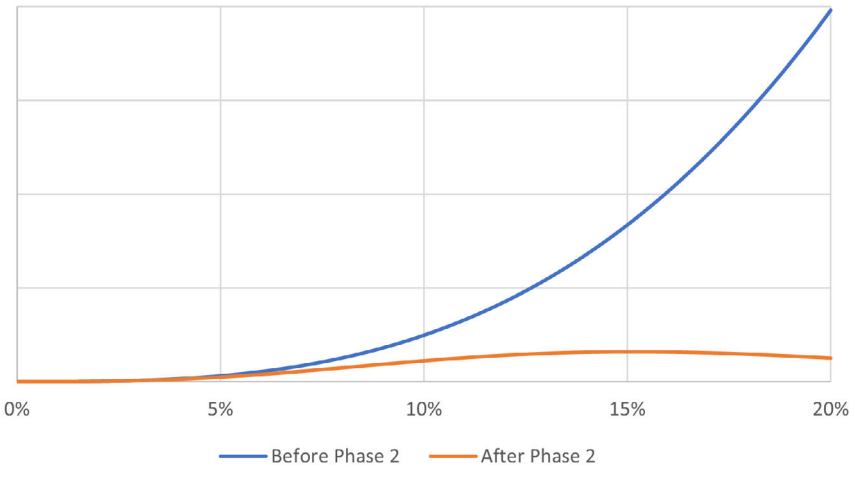

FIGURE 9 Treatment effect distribution before and after Phase 2

fully trusted. The many small trials risk wasting resources. University groups should therefore seek to build consortia rather than launching small, likely inconclusive trials. Government bodies, research councils, etc. should promote collaborative efforts, for example effective platform trials.[20] Major industry and public, sponsors should collaborate, consider the entire program already when designing early trials, utilise group-sequential and adaptive designs to de-risk trials, and factor in selection bias, for example by considering the posterior distribution. Furthermore, with a realistic prior for drug efficacies and given a certain number of patients available for testing in Phase 2 and 3, it would be possible to approach a global optimisation of the testing of the portfolio of all available drugs. That is, one could search for optimal criteria for when to stop development of drugs and re-allocate resources to more promising candidates. Such an approach would have similarities to the concept of parallel drug development as discussed in Wiklund.[21] This could potentially be further developed into a societal decision analysis (cf. Stallard et al.[22]) with the aim of minimising the total number of deaths from COVID-19.

7 | DISCUSSION AND CONCLUSIONS

In the introduction, we outlined the general issue of selection bias, and the problems it may pose to decision makers in correctly assessing the Phase 2 outcomes when making Phase 3 investment decisions. We showed in Ch 3 that the Phase 2 estimate may severely overestimate the true treatment effect, when a positive decision is conditioned on the Phase 2 estimate being statistically significant. This is particularly the case for studies with a small sample size, but the bias can also be substantial for larger trials with small to moderate true treatment effects. This is of importance for decision makers, as the failure to account for the over-estimation may lead to exaggerated claims of treatment effect, eventually implying that futile drug candidates might be taken forward to Phase 3 programmes.

One way of addressing the selection bias in the Phase 2 estimate would be to apply Bayesian methodologies. We derived the posterior distributions for a few selected prior distributions and illustrated some properties of the corresponding posterior mean estimates. For the case that the applied prior distribution is the same as the true treatment effect distribution, the posterior mean is unbiased. This implies that the selection bias would not be an issue if a Bayesian analysis could be performed with the correct prior assumption.[23] In reality the true prior is unknown and the underlying prior assumption will be more or less incorrect. The results presented in Ch. 4 indicate that the choice of prior distribution is of great importance for the ability to reduce the selection bias. As an example, with the true treatment effect following a log-normal distribution, we showed that properties of the Bayesian estimate might be inadequate if the analysis is based on another prior distribution. In particular, applying the mixture normal prior was shown in some situations to enhance, rather than reduce, the selection bias. This is apparent in our results, even though all the compared prior distributions are calibrated to have the same mean, differing only in their shape. The suboptimal performance using an incorrect prior might be suspected to be even worse if also the average of the prior is misspecified. Consequently, the choice of prior distribution for the analysis is important if the correction for selection bias is to be successful. The elicitation of knowledge to inform the choice of prior distribution has been discussed by several authors (e.g., Dallow[24]).

As shown in Ch. 2, the exponential prior leads to a convenient rule of thumb for the adjustment for selection bias, implying that the Phase 2 estimate could simply be adjusted with a multiple of the sample variance of the estimate. This simple rule of thumb was shown to be reasonably successful also in the case where the true prior was not exponential but lognormal. When choosing a prior, one should consider the risk that Phase 2 data are incompatible with the prior. Most importantly, the prior should not give extremely low density for effect sizes that are plausible and close to the Phase 3 go threshold. It is also desirable that the prior does not entirely dismiss the possibility of either surprisingly strong efficacy or negative efficacy. The improper exponential distribution certainly fulfills this. It also has a relatively heavy tail, so that surprisingly good efficacy is not incompatible. It is certainly compatible with negative effects. The fact that it leads to a prior mean even lower than a negative observation is not much of an issue, as a drug with negative efficacy would normally be stopped anyway.

We have in this article illustrated our findings based on a selection of treatment effect (prior) distributions. The selection was made to represent classes of distributions with different properties and to reflect distributions proposed by other authors. We do of course appreciate that the selected distributions merely represent a limited subset, and that the results do not generalise to all potential treatment effect distributions. A similar note could be made regarding the choice of Stop/Go criterion. While we have chosen to illustrate the concepts based on a simple rule defined by

the occurrence of statistical significance, we appreciate that other decision criteria could of course also be relevant (see e.g., Lalonde[25]).

We illustrated in Ch. 5 some empirical findings based on data retrieved from the database at ClinicalTrials.gov. Based on the empirical data we made approximate estimates of the distribution of effect sizes, and the empirical distributions were compared with the prior distributions evaluated in the article. The results indicate that the log-normal and exponential priors appear to be reasonable consistent with the empirical data. As shown in the Supplementary material, the results are relatively consistent across a number of disease areas. The results further indicate that the observed treatment effect is most often smaller than the one anticipated at the planning stage. With the Target Product Profile (TPP) in many cases being overly optimistic, this may lead to an overestimation of the power and an underestimation of the required sample size. Potential consequences of exaggerated TPPs might be that underpowered studies can lead to unnecessary trial failures, partly contributing to the high Phase 2 failure rates, and eventually leading to potentially useful drugs not being developed to benefit patients. We argue the basing study design considerations on reasonable assumptions regarding the treatment effect, and its distribution, is important to avoid such negative consequences.

We do appreciate that the empirical data are not comprehensive, and that there are methodological issues related to its use. We have for instance not been able to identify the intended sign of the estimated treatment effect, and the estimated empirical distribution for the treatment effect does therefore not include negative values. There is also a substantial potential for reporting bias since many trials in the database did not report values for the treatment effect or *p*-value. We also excluded from the analysis trials for which the p-value was not given as a specific value but for instance as $p < 0.05$ or $p > 0.05$. We do however argue that better information on the distribution of the true treatment effect of drug candidate is indeed valuable to guide improved analysis and decision-making and we think that further research in this area is warranted.

While we have used a lot of Bayesian terminology in this article, we appreciate the fact that Bayesian methods may not necessarily be used for the primary analysis of a clinical trial. However, these methods may still be part of the internal planning and decision-making. Hence, an awareness and adjustment for selection bias (potentially by Bayesian methods) would be valuable for the sponsor, even if the formal reporting of the trial would adopt standard frequentist methods. Qu et al.[9] also show how the topic of selection bias can be understood and dealt with from a frequentist perspective.

Selection bias is an important issue. Decision makers and statisticians should be aware of the substantial risk of over-estimating, and consequently the risk of exaggerating efficacy and making false positive decisions. But we also want to emphasise the importance of balancing this issue against the risk of making false negative decisions. As shown by Miller and Burman,[12] it is often beneficial to use much more liberal decision criteria than implied by the standard 5% significance level. These authors show that false negative decisions as a consequence of strict stop/go criteria, may often be detrimental to the expected value of projects. The appropriate choice of decision criteria should not be obfuscated by selection bias considerations.

ACKNOWLEDGMENTS

The authors would like to thank reviewers and editors for valuable comments leading to improvements of this manuscript.

DATA AVAILABILITY STATEMENT

The empirical data used in the article are taken from the publicly available database at www.clincaltrials.gov. The address is also provided in the reference list of the article.

ORCID

Stig Johan Wiklund https://orcid.org/0000-0003-2128-5503

REFERENCES

1. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.
2. Efron B. Tweedie's formula and selection bias. *J Am Stat Assoc*. 2011;106(496):1602-1614.
3. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294(2):218-228.

4. Pereira TV, Horwitz RI, Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. *JAMA*. 2012;308 (16):1676-1684.
5. Chuang-Stein C, Kirby S. The shrinking or disappearing observed treatment effect. *Pharm Stat*. 2014;13:277-280.
6. De Martini D. Adapting by calibration the sample size of a phase III trial on the basis of phase II data. *Pharm Stat*. 2011;10:89-95.
7. Wang SJ, Hung HMJ, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. *Pharm Stat*. 2006;5: 85-97.
8. FDA, US Food & Drug Administration. 22 case studies where phase 2 and phase 3 trials had divergent results. 2017. https://www.fda.gov/downloads/aboutfda/reportsmanualsforms/reports/ucm535780.pdf. .
9. Qu Y, Du Y, Zhang Y, Shen L. Understanding and adjusting the selection bias from a proof-of-concept study to a more confirmatory study. *Stat Med*. 2020;39:4593-4604. https://doi.org/10.1002/sim.8740.
10. Kirby S, Burke J, Chuang-Stein C, Sin C. Discounting phase 2 results when planning phase 3 clinical trials. *Pharm Stat*. 2012;11(5): 373-385.
11. Chuang-Stein C, Kirby S. *Quantitative Decisions in Drug Development*. Cham, Switzerland: Springer Series in Pharmaceutical Statistics; 2017:9783319460765.
12. Miller F, Burman CF. A decision theoretical modeling for Phase III investments and drug licensing. *J Biopharm Stat*. 2018;28:698-721.
13. Wiklund SJ. A modelling framework for improved design and decision-making in drug development. *PLoS ONE*. 2019;14(8):e0220812. https://doi.org/10.1371/journal.pone.0220812.
14. Frewer P, Mitchell P, Watkins C, Matcham J. Decision-making in early clinical drug development. *Pharm Stat*. 2016;15:255-263.
15. ClinicalTrials.gov. https://clinicaltrials.gov/ .
16. Zhou F, Yu T, Du R, Fan G, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395:1054-1062.
17. Richardson S, Hirsch JS, Narasimhan DO, et al. Presenting characteristics, comorbidities, and outcome among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA*. 2020;323:2052-2059.
18. Horwitz LI, Jones SA, Cerfolio RJ, et al. Trends in COVID-19 risk-adjusted mortality rates. *J Hosp Med*. 2020;16:90-92. https://doi.org/10.12788/jhm.3552.
19. Ledford H. Why do COVID death rates seem to be falling? *Nature*. 2020;587:190-192.
20. Collignon O, Burman CF, Posch M, Schiel A. Collaborative platform trials to fight COVID-19: methodological and regulatory considerations for a better societal outcome. *Clin Pharmacol Therapeut*. 2021. https://doi.org/10.1002/cpt.2183.
21. Wiklund SJ. Parallel drug development. *Drug Dev Res*. 2012;73:24-34.
22. Stallard N, Miller F, Day S, et al. Determination of the optimal sample size for a clinical trial accounting for the population size. *Biom J*. 2017;59:609-625.
23. Dawid AP. Selection paradoxes of Bayesian inference. *Multivariate analysis and its applications*. Institute of Mathematical Statistics: Hayward, CA; 1994:211-220. https://doi.org/10.1214/lnms/1215463797, https://projecteuclid.org/euclid.lnms/1215463797.
24. Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. *Pharm Stat*. 2018;17:301-316.
25. Lalonde RL, Kowalski KG, Hutmacher MM, et al. Model-based drug development. *Clin Pharmacol Therapeut*. 2007;82:21-32.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wiklund SJ, Burman C-F. Selection bias, investment decisions and treatment effect distributions. *Pharmaceutical Statistics*. 2021;20(6):1168–1182. https://doi.org/10.1002/pst.2132