UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Strong versus weak prior for predicting patient recruitment at the start of clinical trials in the context of drug availability issues

Auteur: Clément Laloux

Promoteurs: Catherine Legrand & Marco Munda

Lecteur: Vincent Bremhorst

Année académique 2019-2020

Acknowledgment

First of all, I would like to thank my promoters, Catherine Legrand and Marco Munda, for their help and support during the writing of this academic thesis. Their advice and guidance helped me to push the reflections on my research further and further each time.

I would also like to thank Pharmalex and again Marco, which offered me the opportunity to work on this topic and explore it with my own ideas and knowledge.

I also have gratitude to Vincent Bremhorst who was interested in my project and agreed to read it.

I want to give a special thanks to my older brother Timothée, who red and corrected this thesis countless times to advise me. In the end, I am sure he knows it as much I do. I also want to thank my other older brother Thomas, who provided me with great feedback on the last version of this work and Rémi Georges who helped me for the last corrections.

Eventually, I want to thank all the persons that supported me during this thesis, my other relatives and my friends, specifically my library team: Coline and Ambre. Moreover, I want to thank all the LSBA members and professors who taught me and help developed my statistical knowledge during my master here.

Table of contents

List of figures

List of tables

1 Introduction

During clinical trials, subjects' recruitment represents an element of major importance, as it may have several impacts on their successes. In the first stages of a trial, investigators need to determine a sample size, namely, the number of patients to be recruited in order to have adequate power to demonstrate a treatment effect (Jiang & al., 2014). In addition, this choice helps them to make key trial decisions related to the duration, the budget or the drug submission planning. Hence, failing to achieve the target sample on time can affect studies on multiple levels.

One of the main issues is related to the validity of the analyses, since the study will not reach its pre-planned statistical power. It leads to results "*less likely to obtain sufficient precision and make meaningful scientific inferences*" (Zhang & Long, 2010, p. 649; Jiang & al., 2016). It also affects the trial financially in two ways. (i) It can generate increased costs to extend recruitment to meet the required number of patients. (ii) It can generate a financial loss if the trial is terminated and was unable to meet the adequate number of subjects, because with less statistical power, it lowers the trial scientific relevance and provides inconclusive results (Jiang & al., 2016). Two other major consequences of concern may arise. (i) The trial supply chain will no longer be appropriate because with a slower patient recruitment than expected, the planned production and distribution of drugs would not correspond to the real demand. (ii) Since it delays the availability of new therapies and treatments, the medical progress would slow down (Jiang & al., 2014). In this thesis, we will investigate a Bayesian method to help clinical trials improve their monitoring of patient recruitment to be best able to anticipate those issues. The method aims to predict the most accurately possible patient recruitment in ongoing trials.

Over time, it has been observed in clinical trials that researchers tend to underestimate the total recruitment duration for their target sample size. It leads to overoptimistic rates for patient accrual and, thus, causes delays (Jiang & al., 2016). As an example, Sully & al. (2013) looked at a cohort of 72 multicentre trials supported by two UK funding groups (the Medical Research Council and the Health Technology Assessment Programme) between 2002 and 2008. Their main results showed that, out of the 72 trials, fewer than one third (29%) were able to achieve their target sample size within the initially planned duration. Amongst the other trials, 25% succeeded thanks to an extension while 46% simply failed to complete this objective. Similar outcomes were obtained by Pemberton & al. (2018), which investigated the success of 167

cardiovascular trials from 1996 to 2012 (carried out by the National Heart, Lung, and Blood Institute). In that case, only 26.3% of the trials succeeded in achieving their target number of participants within the planned time frame. Overall, many more articles highlight the trialists overconfidence in accrual rates (see Jiang & al., 2014).

Another important issue in the management of clinical trials concerns the drugs availability on medical sites. It is a matter of fact that patients represent the usual bottleneck in the progress of a clinical study, since it cannot go forward without any subject entering in it. When a patient enters a trial, drugs need to be available on-site, otherwise it can have significant consequences on the successful completion of a trial (Fleischhacker & Zao, 2011). A major impact is obviously delays, as subjects will have to wait to start the treatment or worse, could even drop the trial before even starting it. A simple solution would be to anticipate and produce large inventories to be stored in each medical site. However, the problem is more complex, because of several elements:

- Firstly, the storage of drugs involves holding costs, i.e. an estimated fixed cost to converse one unit of a drug in the proper conditions for use. Large inventories can therefore translate into high holding costs for pharmaceutical firms. Moreover, clinical centers may not have the storage capacity to support high volume of supplies required (Peterson & al., 2004).
- Secondly, drugs in clinical trials sometimes involve relatively short shelf life, making it impossible to create large long-term inventories. Indeed, if a firm would like to anticipate the demand for a long duration, some of the drugs may expire before they are used and would thus require resupply (Peterson & al., 2004).
- Thirdly, unused or outdated drugs involve destruction costs for firms, since they must be treated with a specific attention in proper disposal facilities (Fleischhacker & Zao, 2011). As an example, "*Cotherix Inc., estimated $126,000 in destruction costs for an obsolete drug that was valued at $1.5 million*" (Cotherix Inc., 2006 as cited in Fleischhacker & Zao, 2011, p. 497).
- Fourthly, if a large number of sites does not meet expected recruitment, it would generate huge waste for the firms, since "*unused materials cannot be re-routed to other sites due to regulatory restrictions*" (Chen & al., 2010, p. 146).

- Fifthly, drugs for clinical trials generally involve high fixed and variable production costs "*due to the low demand volume, low yield and the premature manufacturing process*" (Frleischhacker & Zao, 2011, p. 497).

As a consequence, regarding drug productions, firms need to find a trade-off between producing large quantities not to incur high fixed costs while avoiding costs related to inventories and waste. Regarding trials overall, firms need to balance between the latter drug-related costs, and the costs incurred by delays in the trial. In other words, they need to ensure that medical centers do not face a shortage while they do not overproduce. Moreover, there are situations where only a limited quantity of drugs is available and no short term production can be envisioned. Thus, trialists need to find the best way to allocate it among centers to avoid shortage.

The problems presented above are of some concern given the role of clinical trials in medical development and stress the importance of developing efficient and reliable quantitative tools to plan and monitor patient recruitment (Gajewski & al., 2008). It would improve investigators' decisions related to sample sizes and duration targets, making them more realistic and attainable. Moreover, they would detect any issue affecting accrual rates early during the enrollment, which would help to conceive appropriate corrective plans or take a decision to halt or continue the trial (Gajewski & al., 2008). Concerning drugs availability, it would help to avoid shortage of drugs in medical centers. Furthermore, it would help companies to produce the most appropriate quantity of drugs to avoid waste and high holding costs.

A major element of those tools involves the ability to model and predict patient accrual accurately, either before or during the recruitment. The prior prediction is of main consideration, as it will help trialists at the design stage. It will give them complete information on plausible patient arrival dates with multiple scenarios (confidence region), preparing them to anticipate any issue concerning enrollment. The ongoing prediction of subjects, for its part, will ensure that everything goes according to the plan throughout the study (Gkioni & al., 2019). To this day, predicting the recruitment remains one of the main concerns for trialists as it was highlighted by Healy & al. (2018).

Several studies have explored methods and models to predict recruitment. A comprehensive review is available in Gkioni & al. (2019). Those models generally focus on the prediction at the design stage of trials. Some of them also allow a continuous monitoring, generally by using a Bayesian approach. Its advantage consists in (i) the construction of an informative prior to

obtain expected recruitment at the design stage based on the investigators knowledge or on data from similar previous trials. (ii) During the enrollment, it combines these prior beliefs with the observed recruitment data in order to update the prediction, namely posterior prediction, to guarantee that the trial is still on the right track (Gkioni & al., 2019).

In this thesis, we study this second advantage of the Bayesian approach in the context of drug availability problems in clinical trials. As we mentioned, predictions have a crucial role in the conception of an appropriate drug distribution planning for medical centers, since they provide the information on the drug quantities needed in each period. Therefore, methods used to compute them need to be the most accurate possible, to prevent centers from facing a shortage. If well-established methods already exist for the monitoring of an ongoing trial (see Barnad et al, 2010; Gkioni & al., 2019), some issues still arise regarding the predictions at the beginning of a trial when few data are available (Gajewski & al, 2008; Zhang & Long, 2010; Jiang & al, 2014). The goal of this work is, then, to explore an approach to answer this problem.

The main issue faced by trialists at the beginning of trials is the limited guidance provided by the data. As the number of subjects is small, there is a lot of uncertainty around predictions, which, in return, affects the ability of trialists to derive meaningful information on future recruitment (Zhang & Long, 2010). One of the advantages of Bayesian methods is that it allows investigators to provide information into a model with a certain precision, i.e. the prior, which is combined with data. The more precise (strong) it is, the more it lowers uncertainty around predictions. Hence, using a valuable strong prior information into a model when the amount of data is limited can have positive impacts on a trial, as it would, for example, allow trialists to make a more efficient drug distribution planning, meaning fewer waste and drug shortages in medical centers (Zhang & Long, 2010). However, the problem of strong priors lies in their inability to react quickly in cases of misleading prior information, i.e. a prior information which differs from what the data show (Jiang & al., 2014).

Knowing that many trials failed to be completed on time due to overconfidence in the expected accrual rate, several questions arise regarding the use of strong prior information on the predictions of recruitment for a trial:

- What would be the consequences of a strong informative prior that is significantly off target?
- What are the benefits of a strong informative prior that is significantly on target?

- For both questions, what are the differences if a weak prior is used instead?

Exploring the benefits and consequences of the information provided in the prior on the model appears truly important in the context of drug distribution planning. Therefore, the objectives of this thesis are twofold: (i) determine the most relevant level of information to introduce into the modeling of ongoing trial recruitment and (ii) evaluate the relevance of introducing strong prior information in the model. With this thesis, we aim at being able to give relevant advice to trialists for monitoring stage-related decisions. It is important to note that, in this work, we are not interested in the construction of the most relevant prior, nor assessing the influence of different types of prior on the predictions. Our only interest lies in studying the influence of information levels in the prior.

This work is based on a statistical method first developed and introduced by Gajewski & al. (2008). It is currently implemented successfully in the business by Pharmalex, a consulting company specialized in pharma, biotech and medical device sectors. The method aims at modeling the recruitment data of an ongoing trial to predict future patients' arrivals. It takes a time-to-event approach to model the waiting times, i.e. difference in date between two arrivals. Moreover, by using the Bayesian framework, it accounts for uncertainty in the predictions of recruitment by providing a predictive distribution from which different metrics can be derived, such as credible intervals.

The first chapter of this thesis aims at providing all the required knowledge to understand the thesis. First, we will introduce the main characteristics of the Bayesian statistics as well as the Markov Chain Monte Carlo approach. Both represent essential elements in the method to predict recruitment. In the second chapter, we will describe with more details the objectives of this thesis and the methodology applied. Then, we will present the results of the simulations studies, which constitutes the core of this thesis, as it represents the basis to explore our objectives. In the third chapter, we will sum up the main findings of the simulation studies, compare them to the scientific literature on the topic, present their limitations, discuss their main implications for clinical trials and, eventually, provide some areas of improvement. In the last chapter, we give our main conclusions on the thesis.

2 Theoretical context

2.1 Bayesian Approach

One of the main aspects of statistics lies in the statistical inference, which, based on observable data, allows general conclusions to be drawn about unobserved quantities. For example, a pharmaceutical firm wants to know the efficiency of a new drug compared to the one currently used to treat a specific disease. To make the comparison, they need to test the effect of both drugs on people affected by the disease. However, it is impossible to visit every sick person. Instead, statisticians have to rely on samples of the population of interest. Thus, the comparison between the two drugs has to be done on a sample of affected people, with the objective to conclude about the (true) treatment effect in the population of affected people. Statistical inference provides methods to achieve this (Gelman & al., 2013).

Bayesian statistics provide methods to make statistical inference. In the Bayesian approach, parameters of interest are considered as random quantities and, like data, possess a distribution. In the first place, this distribution represents all the knowledge someone has on a parameter of interest and is referred as the prior distribution or simply the prior. Then, by adding a new source of information, i.e. relevant data, this prior knowledge will be updated, resulting in the posterior distribution (King, 2010). The Bayesian inference can be seen as "*reallocation of credibility across possibilities*" (Kruschke, 2015, p.22). A first proposal of possible values (possibilities) for the parameter of interest is defined, as well as their probabilities (also referred as credibilities), the prior. Then, each time data is added, these probabilities change. Values that are more consistent with data are assigned higher probabilities, whereas those that are not are assigned smaller probabilities (Kruschke, 2015). In the end, the posterior distribution provides a complete information on a parameter of interest, giving the set of all possibilities together with their credibility.

Formally, Bayesian data analysis can be divided into several steps (Kruschke, 2015, p.25):

1. Identify the relevant data to the study.
2. Choose a statistical model that fits the relevant data. "*The mathematical form and its parameters should be meaningful and appropriate to the theoretical purposes of the analysis*" (Kruschke, 2015, p.25).
3. Define a prior distribution for the parameter(s) of the model. It must represent the subjective beliefs about likely values of the unobserved parameters, before any data is

collected.

4. Use Bayesian inference to update the distribution of the parameters and derive from it the quantities of interest relevant for the study.
5. Perform a model check to evaluate how well it fits the data. If it does not seem relevant, consider a different descriptive model and repeat the steps.

In this part, we will not elaborate on steps 1. and 5., as they will not be addressed in this work. However, the remaining steps may require further explanations, by introducing some basic concepts and notations of Bayesian inference.

In this thesis, only the case of a univariate continuous distribution with a single unknown parameter is considered. So in the next section, D corresponds to n x 1 vector of observed data. λ denotes some unobserved parameter of a statistical model. f(·) denotes a marginal density distribution, with the arguments determined by the context. f(·|·) denotes a conditional density distribution.

2.1.1 Bayes' Theorem

As introduced, Bayesian inference starts with some prior beliefs on a parameter of interest and then updates them with relevant data to the study. Thus, a way to combine those two quantities is necessary. The Bayes' theorem is the mathematical relation which connects the information of the prior and the data to obtain the posterior distribution of a parameter (King, 2010):

$$f(\lambda|D) = \frac{f(D|\lambda)f(\lambda)}{f(D)} = \frac{f(D,\lambda)}{f(D)} \quad (1)$$

f(λ|D) corresponds to the posterior distribution of λ conditioned on the data, D. The numerator represents the joint probability distribution for both D and the parameter λ, f(D,λ). It includes the statistical model or the distribution assigned to the data (step 2.), also known as the likelihood f(D|λ), as well as the prior distribution elicited for the parameter of interest f(λ) (step 3.). The denominator is the normalizing constant, it ensures the posterior distribution is a proper distribution (i.e. it integrates to 1) (King, 2010). It can be defined by "*the overall probability of the data according to the model, determined by averaging across all possible parameter values weighted by the strength of belief in those parameter values*" (Kruschke, 2015, p.106-107). Formally, it is:

$$f(D) = \int f(\lambda)f(D|\lambda)d\lambda \quad (2)$$

However, note that typically, f(D) is omitted from the Bayes' Theorem, as it is considered a constant independent from the parameter of interest. It yields the equivalent form:

$$f(\lambda|D) \propto f(D|\lambda)f(\lambda) = f(D,\lambda) \quad (3)$$

As explained, f(D) is the normalizing constant "*that makes the area under the posterior distribution sum (i.e., integrate) to one. Dividing by a constant does not change the basic shape of the posterior distribution*" (Ender, 2010, p. 170–171). Thus, the posterior final form is proportional to the prior distribution times the likelihood.

2.1.2 Prior distribution

One of the crucial steps of the Bayesian data analysis lies in the elicitation of a prior distribution for the parameter studied, λ. As we mentioned, it is supposed to represent one's current beliefs on the possible values of λ, before looking at the data. Hence, the form of the density must match the relative probabilities given to each possibility (Boldstad, 2007). This choice can have a major role on the inference results. For example, with sparse data, the prior can narrow the set of λ plausible values to concentrate computations on a specific parameter space. On the contrary, it can enable the computation to take place on values that would otherwise be impossible given a specific data distribution (King, 2010).

When choosing a prior, two main situations are encountered (King, 2010):

- No prior information on the parameter of interest is available.
- It exists some prior knowledge on the parameter of interest that can be expressed into a suitable distribution.

In both cases, several methods can be used to specify the prior distribution. They are generally situation-specific. In this thesis, only one method will be considered, which involves a subjective input. This particular type of prior that is often used for its convenience when assessing product of f(D|λ) and f(λ) is the conjugate prior (King, 2010).

Conjugate priors represent an important aspect of Bayesian statistics. They correspond to situations "*when a posterior distribution for a model parameter has the same distributional form as the corresponding prior*" (King, 2010, 76). In practice, conjugate priors only exist for parameters of distributions in the exponential family. Exponential family provide the advantage of having conjugate priors which have the same functional forms as the likelihood functions. It results in posterior distributions that retain those functional forms, and, thus eases their

calculations. Indeed, no integration, or simple ones, are necessary. One just needs to update parameters of the prior with the observations to obtain posterior values (Boldstad, 2007). This is illustrated in **Table 1.**, that displays examples of priors conjugated to likelihood functions and gives their subsequent posterior results.

f(D|λ)	f(λ)	f(λ|D)
Normal $N(\lambda, \sigma^2)$	Normal $N(\mu, \tau^2)$	$N(\rho(\sigma^2\mu + \tau^2 \sum D), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $P(\lambda)$	Gamma $G(\alpha, \beta)$	$G(\alpha + \sum D, \beta + 1)$
Gamma $G(\nu, \lambda)$	Gamma $G(\alpha, \beta)$	$G(\alpha + \nu, \beta + \sum D)$
Binomial $B(n, \lambda)$	Beta $Be(\alpha, \beta)$	$Be(\alpha + D, \beta + n - \sum D)$
Negative Binomial $Neg(m, \lambda)$	Beta $Be(\alpha, \beta)$	$Be(\alpha + m, \beta + \sum D)$
Multinomial $M_k(\lambda_1, \dots, \lambda_k)$	Dirichlet $D(\alpha_1, \dots, \alpha_k)$	$D(\alpha_1 + \sum D_1, \dots, \alpha_k + \sum D_k)$
Normal $N(\mu, 1/\lambda)$	Gamma $G(\alpha, \beta)$	$G(\alpha + 0.5, \beta + \left(\mu - \sum D\right)^2 /2)$

Table 1. Posterior distributions for some conjugate priors from the exponential family*. It is assumed that all parameters except λ are known. ∑ D correspond to the sum of all the observed data (Robert, 2001, p.121).*

As mentioned, the use of a conjugate prior automatically induces information in the model. Indeed, the distribution elicited provides details about central tendencies as well as measures of spread of the parameter of interest. However, one can make vary the amount of information introduced in the prior so that a high uncertainty is placed around the parameter of interest central value. The latter is referred as a weak prior information. Gelman & al. (2013, p.55) defines it as "*a proper distribution but set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available*". Only a limited input is included to make sure the posterior distribution is meaningful according to a statistician's prior beliefs (Gelman & al., 2013).

2.1.3 The posterior is a compromise between the prior and the likelihood

Following the second equation of the **Bayes' Theorem** section, only the prior distribution and the likelihood must be evaluated to obtain the posterior distribution. Thus, its computation obviously represents a compromise between the information provided by those two quantities. In other words, the posterior can be seen as a weighted average of prior information on λ and information provided by the data. Therefore, it is important to explain how those weights work, how they evolve throughout the process and what are the theoretical influences of the sample size and the prior (Kruschke, 2015).

Considering the sample size, a simple and intuitive influence is that a larger quantity of data lead to a lower uncertainty around possible values of the parameter of interest. For example, considering the case of a conjugate prior, a large sample size translates into a decrease in the spread of the posterior distribution (Kruschke, 2015). Regarding its impact relative to the prior, the same kind of reasoning can be applied, an increasing sample size leads to a smaller influence of the prior. It represents a general characteristic of Bayesian approach that "*the compromise is controlled to a greater extent by the data as the sample size increases*" (Gelman & al. 2013, p.32). In case of very large datasets, one might even observe that quite different priors would give almost the exact same posterior distribution, because the amount of data would dominate and wipe out any prior influence (Boldstad, 2007). However, the sample size necessary to shift influence will depend on the prior distribution, i.e. the uncertainty placed around the prior beliefs on λ.

Table 2. illustrates this combination between the data and the prior for some conjugate priors in terms of a central tendency measure, the mean. In this case, it is obvious that the more data in the model, the more it will influence the mean, since prior parameters remain fixed, while sample sizes keep on increasing.

f(D|λ)	f(λ)	Posterior mean
Normal $N(\lambda, \sigma^2)$	Normal $N(\mu, \tau^2)$	$\frac{\sigma^2\mu + \tau^2 \sum D}{\sigma^2 + \tau^2}$
Poisson $P(\lambda)$	Gamma $G(\alpha, \beta)$	$\frac{\alpha + \sum D}{\beta + 1}$
Gamma $G(\nu, \lambda)$	Gamma $G(\alpha, \beta)$	$\frac{\alpha + \nu}{\beta + \sum D}$
Binomial $B(n, \lambda)$	Beta $Be(\alpha, \beta)$	$\frac{\alpha + \sum D}{\alpha + \beta + n}$
Negative Binomial $Neg(m, \lambda)$	Beta $Be(\alpha, \beta)$	$\frac{\alpha + m}{\alpha + \beta + \sum D + m}$
Multinomial $M_k(\lambda_1, \ldots, \lambda_k)$	Dirichlet $D(\alpha_1, \ldots, \alpha_k)$	$\frac{\alpha_i + \sum D_i}{\left(\sum_j \alpha_j\right) + n}$
Normal $N(\mu, 1/\lambda)$	Gamma $G(\alpha, \beta)$	$\frac{\alpha + 1}{\beta + (\mu - \sum D)^2}$

Table 2. Posterior distribution means for some conjugate priors from the exponential family*. It is assumed that all parameters except λ are known. ∑D correspond to the sum of all the observed data (Robert, 2001, p.176).*

Regarding the prior, its influence depends on the previous knowledge one has on the parameter λ, and, of course, its level of confidence about it. When a vague prior belief is used, or a weak prior information in this work, the likelihood will dominate the posterior, since it contains more information about the parameter of interest. Moreover, in the limiting case where no prior information is available at all, the posterior will reasonably only be determined by the data (Jackman, 2009). Nevertheless, when information is provided to the model, it takes more data to shift the influence towards the likelihood. Indeed, with accurate and meaningful prior beliefs, it seems like a logic reasoning, "*because in real applications a sharp prior has been informed by genuine prior knowledge that we would be reluctant to move away from without a lot of contrary data*" (Kruschke, 2015, p.114). Besides, in the limiting case of a unique prior value with little or no variability, the posterior will reasonably only be determined by the prior, likelihood being completely ignored (Jackman, 2009).

Overall, Bayesian reasoning is intuitively rational when considering the compromise between information from the prior and the data. When the knowledge about a parameter of interest is

limited or nonexistent (probabilities spread over a lot of possibilities), one prefers to let the data "talk" and provide the information in the model. However, when a valuable knowledge about this parameter exists, which can be translated into a relevant and precise prior information (all probabilities focus on a narrow space), then it takes a large amount of data to shift those prior beliefs (Kruschke, 2015). In the end, whatever the prior, "*the role of new data is to modify the beliefs from whatever they were without the new data*" (Kruschke, 2015, p.114).

This process is illustrated in **Figure 1**. It displays the two situations expressed in the latter paragraph. The left part of the graph shows that with a weak prior and few data, the posterior will follow the likelihood. The right part shows that with strong prior and a larger sample size, the posterior is still significantly influenced by the prior, since the mode is still highly leaning on the prior one. Another remark is that, as we mentioned, with more data in the analysis, we obviously obtain more certainty about parameter possible values. Indeed, the interval computed with 4 observations is much larger than the one with 40 (Kruschke, 2015).

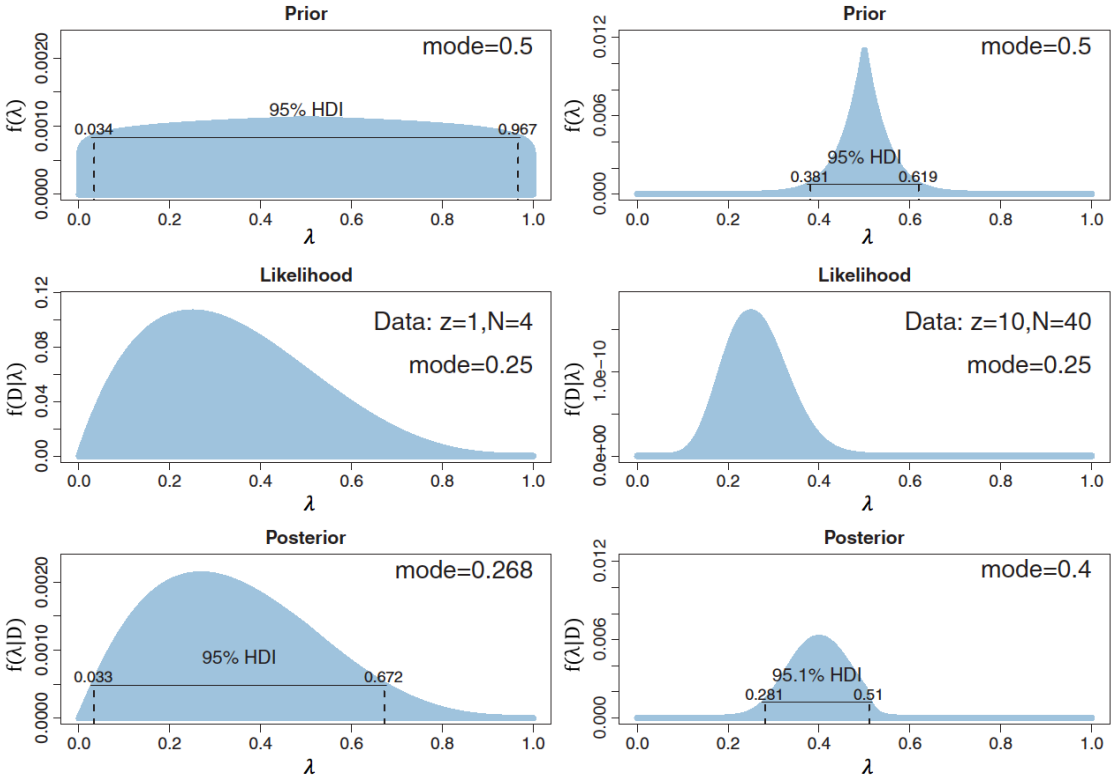

Figure 1. Priors, likelihoods and posterior densities for parameter λ in [0,1]. *The left part presents a combination of a flat prior and a likelihood with only 4 data. The right part presents a combination of an informative prior and a likelihood with 40 data. In each case, the mode is given. (Kruschke, 2015, p.114).*

2.1.4 Bayesian quantity of interest

After having calculated the posterior distribution, we might be interested in several quantities in the Bayesian analysis. The first quantity is the marginal posterior distribution of the parameter studied λ. Obviously, in a one unknown parameter setting, it is directly given by the posterior distribution, thus all the information about it is already available above.

Another quantity of interest in Bayesian inference involves the ability to deal with prediction problems, i.e. be able to make predictions about unknown future observations of the relevant data to the study. The method to achieve it has a similar logic to what has been presented so far (Gelman & al., 2013). To be more precise, it consists in deriving the "predictive distribution" of the data, which depends on two quantities, the posterior on λ and the likelihood (Gelman & al., 2013, p.7):

$$\begin{aligned} f(D_{pred}|D) &= \int f(D_{pred}, \lambda|D)d\lambda \\ &= \int f\left(D_{pred}\middle|\lambda, D\right)f(\lambda|D)d\lambda \\ &= \int f\left(D_{pred}\middle|\lambda\right)f(\lambda|D)d\lambda \end{aligned} \tag{4}$$

Hence, the predictive distribution is conditional on the observed data. Note that the last step is based on the conditional independence of D and D_{pred} given λ (Gelman & al., 2013). Then, to generate predictions, one must draw a random sample from the predictive distribution.

Alternatively, one may consider another method than the analytical one to predict data points. Based on the information from the posterior predictive distribution formula, a simple and easily generalizable algorithm consists in generating a random value λ_1 from $f(\lambda|D)$, then generate a random vector of predictions $D_{pred,1}$, …, $D_{pred,\ n}$ from the statistical model $f(D_{pred}|\lambda)$ with parameter λ_1 (Gajewski & al., 2008). This approach yields equivalent results than analytical solutions and, moreover, enables computation when no close form on the posterior predictive is integrable.

2.1.5 Credible Intervals (CI)

Probability distribution contains all the information about a specific variable or a parameter of interest. For example, in a Bayesian context, the posterior distribution would represent all the

current knowledge one has on a specific parameter. However, this information can be wide and sometimes not directly interpretable, thus, it may be interesting to summarize its content. Several measures exist to achieve this, all with their own interpretation. One might want to know about central tendencies of the distribution, commonly reported by mean, median or mode or about the spread of the values, commonly reported by the variance or the interquartile range (Gelman & al., 2013).

A valuable information when studying a parameter of interest or a predictive distribution is their uncertainty, i.e. whether one can be confident about a specific value for instance. In Bayesian statistics, a relevant description of uncertainty is provided by credible intervals (CI). Bayesian credible intervals give "*a range of values that has a known high probability (1-α), of containing the true value of the parameter*" or any value of interest (Boldstad, 2007, p.140). To illustrate this, suppose we are interested in a parameter λ with posterior distribution f(λ|D) (where D represents the data). Then, a (1-α) credible interval with a lower bound λ_{lb} and an upper bound λ_{ub} would be defined as (King, 2010, p.86):

$$\int_{\lambda_{lb}}^{\lambda_{ub}} f(\lambda|D)d\lambda = (1-\alpha) \; where \; 0 \leq \alpha \leq 1 \tag{5}$$

The latter definition does not lead to a single solution. Thus, different intervals are derivable for the same distribution. A classic interval, somewhat analogous to confidence intervals approach, consists in symmetric (1-α) credible interval. It is made so that the areas outside the interval represent equal parts, each corresponding to α/2 (King, 2010, p.86):

$$\int_{-\infty}^{\lambda_{lb}} f(\lambda|D)d\lambda = \frac{\alpha}{2} = \int_{\lambda_{ub}}^{+\infty} f(\lambda|D)d\lambda \tag{6}$$

Another popular way to compute Bayesian credible intervals is the highest density interval, referred as HDI. It is made such that all the values of the distribution inside the interval (1-α) "*have a higher credibility than any point outside the interval*" (Kruschke, 2015, p.87). Thus, HDI provides the most credible values for a parameter or a variable of interest which covers most of the distribution. Formally, HDI holds two main characteristics. Firstly, it contains (1-α)% of the total probability mass. Secondly, it is constructed such that the lower and upper bounds of the intervals have equal heights of the density (Boldstad, 2007; Kruschke, 2015), which gives the following definition:

$$\int_{\lambda_{lb}}^{\lambda_{ub}} f(\lambda|D)d\lambda = (1-\alpha)\ with\ f(\lambda_{lb}|D) = f(\lambda_{ub}|D) \quad (7)$$

As a consequence of the second characteristic, the areas outside the interval do not necessarily have equal sizes. This is shown in **Figure 2**, where HDI clearly does not match the symmetric intervals. Moreover, it represents the shortest interval containing (1-α)% of the λ density.

Regarding differences between the two methods, it depends on the form of the distribution analyzed. When a distribution is symmetric and unimodal, they both lead to the same intervals. The differences emerge when facing multimodal or skewed distributions as shown in **Figure 2**, with intervals varying in width and location. In practice, the symmetric intervals will be reported, since it is easy to compute (Gelman & al., 2013). However, what about the relevance of the information communicated on the parameter of interest? In the **Figure 2**, several values of λ outside the symmetric interval (on the left) are better supported than some others within the interval (on the right), while, for HDI, it does not happen, all values inside have a higher probability density than any value outside. In the case of a skewed distribution, the interval providing the most relevant information about λ is the HDI, because it has a shorter width and it provides the most credible values about λ (Kruschke, 2015). As a consequence, in this thesis, the HDI will be the method used to compute credible intervals, as we do not know the kind of distributions analyzed.

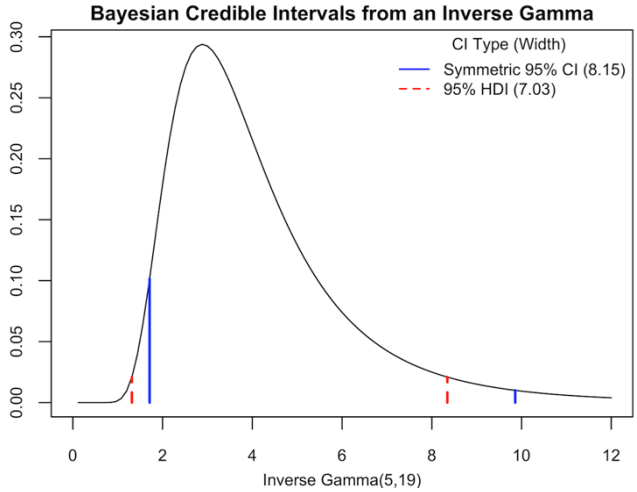

***Figure 2. Bayesian Credible Intervals comparison**. It shows the symmetric CI and the HDI in the case of an Inverse Gamma distribution, which is skewed. CI widths are reported in the legend.*

2.2 Markov Chain Monte Carlo

Thus far, we have only introduced an analytical method using conjugate prior distributions to obtain the posterior distribution f(λ|D) (see **Bayes' Theorem** and **Prior distribution**). In the latter, the posterior is computed in closed form without having to solve complicate integrals (Gelman & al., 2013). Nevertheless, this method is not applicable in every situation. Indeed, in some studies, the prior expressing one's knowledge cannot be translated into a conjugate prior. The finding of an analytical solution in those situations becomes impossible. Therefore, efficient methods have been developed to approximate numerically the posterior distribution and derive their main quantities of interest (Kruschke, 2015). This is referred by King & al. (2010, p.69) as "*the modern Bayesian approach analysis, which does not try to integrate the posterior joint distribution analytically, but instead to employ special simulation procedures which result in samples from the posterior distribution*". One of the most common approach is known as Markov Chain Monte Carlo (MCMC). It can be divided into two parts as described below, Monte Carlo integration, to derive the main quantities of interest, and Markov chains, to compute the approximation of the posterior distribution (King & al., 2010).

2.2.1 Monte Carlo Integration

The Monte Carlo integration consists in providing a solution when facing complex integrals that cannot be calculated explicitly. In a Bayesian context, it is very common to not simply assess the posterior distribution but to summarize it using descriptive statistics which require integration on the density, like the mean or the median. However, that integration is not always feasible (King & al., 2010). For example, one is interested in the posterior expected value of a function ψ(·) of a parameter of interest λ given the data D, the following integral must be evaluated (King & al., 2010, p.99):

$$E\big(\psi(\lambda)\big) = \int \psi(\lambda) f(\lambda|D) d\lambda \tag{8}$$

Monte Carlo provides a numerical alternative to approximate this result by using a simulation approach. The method generates a random sample from the posterior distribution of λ, then evaluates the empirical estimate of the expectation (King & al., 2010). For instance, if we draw t observations $\lambda_1, \ldots, \lambda_t$ from f(λ|D), the estimate of the integral is (King & al., 2010, p.99):

$$\bar{\psi}_t = \frac{1}{n} \sum_{i=1}^{t} \psi(\lambda_i) \tag{9}$$

Therefore, in this specific case, one just needs to compute the sample mean to approximate the result. Moreover, given that the generated sample is independent, the law of large numbers ensures the convergence in probability for the sample mean (King & al., 2010).

In practice, this method can be applied to any quantity of interest related to the posterior, given that we use a large independent sample. In the event that someone wants to know the probability that λ is lower than a specific value *a*, it could simply draw a sample from the posterior, then compute the proportion of values lower than *a*. As we mentioned, this method is really useful when integrals are complex, as well as when the density of $f(\lambda|D)$ is not easily definable, i.e. there is no analytical form (King & al., 2010).

Overall, the main issue encountered using Monte Carlo integration consists in the ability to generate large and independent samples from complex posterior distributions where no analytical solution is available. A consistent and reliable method is required to achieve this. Fortunately, it is precisely the purpose of Markov chains presented in the next section (King & al., 2010).

2.2.2 Markov chains

The Markov chains is a general approach to compute samples from arbitrary posterior distributions. It aims at approximating this distribution by generating values of a parameter of interest sequentially. At each step, the generated value λ_t is obtained by means of a transition distribution $T_t(\lambda_t|\lambda_{t-1})$ that only depends upon the previous value λ_{t-1}, starting from a chosen initial value λ_0 (Gelman & al., 2013). "*The key to the method's success is that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution*" (Gelman & al., 2013, p. 275). In the end, the approach provides a sample (a chain) $\lambda_1, \ldots, \lambda_t$ approximating the target posterior distribution $f(\lambda|D)$. Note that Markov chains just represent the overall process introduced here. There exist several methods to apply it, including the one used in this thesis, which will be presented later in this part.

Concerning the sample independence, it is clear that some dependence between the λ_i are induced, because each value of λ depends on the previous one. However, some manipulations can be done to obtain a state of quasi-independence, allowing the use of Monte Carlo integration

on the generated samples. A first manipulation consists in running several chains simultaneously, such as no chain has any impact on another one. Thus, we only get dependence between the λ_i from the same chain. A second manipulation relies on batch sampling. In that case, instead of simply taking all the values in the chain, only one is taken for the future sample every *c* iteration. Hence, the objective is to keep members of the chain which are far enough apart, so that they have as little effect as possible on other values in the sample or, in other words, to lower their dependencies in the sample (Robert, 2001). Obviously those two manipulations can be combined to enhance the quasi-independence state.

2.2.3 MCMC approach

The Markov Chain Monte Carlo approach is the combination of Markov chains with Monte Carlo integration. As we explained, the Markov chains approach generates the sample that converges to the posterior distribution of interest, while the Monte Carlo integration is later used to compute empirical estimates of any posterior summaries of interest (King & al., 2010). This approach does not require the resolution of any integrals, either to obtain the posterior, or to summarize it. Furthermore, it applies to "*(i) continuous and discrete values, (ii) on any number of dimensions, and (iii) with more general proposal distributions*" (Kruschke, 2015, p. 156). Therefore, MCMC allows to work with more complicated models, where analytical solutions do not explicitly exist.

2.2.4 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is an extension of the Metropolis algorithm, a method to generate a Markov chains. To understand it, we must first introduce briefly the Metropolis algorithm. It consists in visiting randomly any value in the parameter space, and only keeping the ones with a relatively high posterior probability (Kruschke, 2015). It works with an acceptance/rejection rule, following the next steps to be achieved (Gelman & al., 2013, p. 278):

1. For the parameter of the model, a starting value λ_0 is chosen. It can, for example, come from its prior distribution, being any summary statistics considered relevant. Note that this value must guarantee $P(\lambda_0|D) > 0$.
2. For t=1, 2…:
    a. At each iteration, we will update the value for λ: $\lambda_{new} = \lambda_{t-1} + z$, with z, a random value from a symmetrical distribution of our choice and λ_{t-1}, the parameter value from the previous iteration.

b. Then we compute the ratio of the density, or the acceptance probability r, specifically expressed as:

$$r = \min\left(1, \frac{f(\lambda_{new}|D)}{f(\lambda_{t-1}|D)}\right) \quad (10)$$

c. We generate a random u ~ Uniform (0,1). If u < r, then accept proposal λ_{new} as λ_t, $\lambda_t=\lambda_{new}$. Otherwise, reject the new value and keep λ_{t-1}, $\lambda_t=\lambda_{t-1}$.

Thus, the Metropolis requires the prior and the likelihood to be easily computed for any given value of λ and any data D. In other words, the values of f(λ) and f(D|λ) need to be easily determined to obtain the posterior value and calculate r (Kruschke, 2015).

The Hamiltonian Monte Carlo differs from the Metropolis in its way to update the value for the parameter of interest at each step. The Metropolis simply generates it from a symmetric distribution centered on the current value of λ, computed at t-1. Hence, the proposal is equally likely to be lower or larger than λ_{t-1} (Kruschke, 2015). In the HMC, the latter sentence does not apply anymore, because proposals are no longer random walk. Indeed, by computing the direction where the posterior distribution increases, referred as its gradient, HMC tends to generate proposals that lean toward values with higher densities than the current one. Thus, the updated value is no more equally likely to be lower or larger than λ_{t-1}. In this way, the algorithm moves much faster through the target distribution (Kruschke, 2015).

A relevant analogy when considering HMC is provided by Kruschke (2015, p. 401) who compares it to "*rolling a marble on the posterior distribution turned upside down*". In theory, in a unimodal distribution, the marble would automatically go to the lowest point. In practice, this is quite different as HMC involves some randomness in the method. The generation of a new proposal could be seen as "*flicking the marble in a random direction and letting it roll around for a certain duration (...) and the magnitude of the flick is randomly sampled from a zero-mean Gaussian*" (Kruschke, 2015, p. 401). When the rolling is stopped, after a predetermined time, the position of the marble represents the new proposal. The idea we want to emphasise here is that the ball can go in any direction, but will favour positions located downhill, in regions of higher posterior probability (Kruschke, 2015).

The flick mentioned above generates a momentum to the marble, which is represented by a variable φ in the algorithm. More precisely, the momentum φ is drawn from a specific

distribution, usually a normal centered on 0 with pre-specified variance (univariate case). Then, for each iteration of the algorithm, the values of φ and λ are updated several times to induce different movements and positions of the ball, last one being the final proposal for λ. The new proposal of λ is thus largely determined by φ (Gelman & al., 2013). Moreover, the distribution of φ is taken into account when computing the acceptance/rejection rule step of the Metropolis, which leads to use the joint distribution f(λ|D) f(φ) (Gelman & al., 2013). However, note that only the parameter studied is of interest, "*the vector φ is an auxiliary variable, introduced only to enable the algorithm to move faster through the parameter space*" (Gelman & al., 2013, p. 301).

Formally, each iteration in the HMC has the following three steps, starting from an initial value λ_0 set by the user like in the Metropolis (Gelman & al., 2013, p. 301-302):

1. Update the momentum φ_{start} by drawing a value from the normal distribution assigned, $\varphi_{start} \sim Normal(0, \sigma_\varphi^2)$.
2. Update simultaneously the momentum and the parameters values several times as describe above. This is done by using a Leapfrog[1] algorithm with L iterations scaled by a factor[2] ε. Like in the Metropolis, values are updated in relation to the other. The Leapfrog steps are:
    a. Compute a new momentum value based on the gradient of the log-posterior density of λ. Divide by 2, to only make a half step of φ:

$$\varphi_{new} = \varphi_{start} + \frac{1}{2}\varepsilon \frac{d \log f(\lambda_{t-1}|D)}{d\lambda} \tag{11}$$

    b. Use the new momentum value to compute a proposal for the parameter of interest:

$$\lambda_{new} = \lambda_{t-1} + \frac{\varepsilon \varphi_{new}}{\sigma_\varphi^2} \tag{12}$$

    c. Complete the update of φ by making another half step toward the gradient value:

[1] "*This algorithm (called a 'leapfrog' because of the splitting of the momentum updates into half steps) is a discrete approximation to physical Hamiltonian dynamics in which both position and momentum evolve in continuous time*" (Gelman & al., 2013, p. 302).

[2] The scale factor ε corresponds to the step size. In other words, it will determine how fast we move far from the original position of the marble. A small ε implies that it will take more steps to move far from initial position, while a large one will obviously allow the marble to move away more quickly (Kruschke, 2015).

$$\varphi_{new} = \varphi_{new} + \frac{1}{2}\varepsilon \frac{d \log f(\lambda_{new}|D)}{d\lambda} \quad (13)$$

After L iterations, the last proposal corresponds to the update on λ.

3. Then we compute the ratio of the density, or the acceptance probability r, including this time the distribution of the momentum evaluated in φ_{start} and φ_{new}:

$$r = \min\left(1, \frac{f(\lambda_{new}|D)f(\varphi_{new})}{f(\lambda_{t-1}|D)f(\varphi_{start})}\right) \quad (14)$$

4. Eventually, generate a random u ~ Uniform (0,1). If u < r, then accept proposal λ_{new} as λ_t, $\lambda_t=\lambda_{new}$. Otherwise, reject the new value and keep λ_{t-1}, $\lambda_t=\lambda_{t-1}$.

Therefore, HMC also requires the prior and the likelihood to be easily determined, as well as being able to compute the gradient of the posterior and, of course, the momentum density for any given values of λ and φ and any data D (Gelman & al., 2013).

3 Assessing the performances of strong versus weak prior

3.1 Objectives

As stated in the introduction, we were interested in the modeling and the prediction of subjects' recruitment in the context of clinical trials. More specifically, we wanted to focus on drug planning issues that can be faced by medical centers. This represents an element of major importance for trials, since inadequate drug planning can have several consequences. For example, it can cause delay if the treatment is not available when a patient arrives in the study or it can generate higher costs for a firm if drugs are overproduced, because of holding and destruction costs. This stresses the importance of developing accurate and reliable prediction methods to help build more efficient drug planning, which, in return would decrease drug shortage and waste. To achieve it, we examined the case of a specific prediction method involving a Bayesian approach.

As explained in **Bayesian Approach**, Bayesian methods offer the opportunity to introduce some prior information, with a certain precision, into a model. At the design stage of a trial, before recruitment has started, the predictive distribution uses only the prior distribution. Thus, in order to obtain the best recruitment prediction, investigators must put the most accurate information available on the prior, including all the previous experiences obtained from similar trials. However, during the monitoring stage, with subjects arriving in the trial, a new source of information is introduced and the posterior prediction becomes a mix between the observed data and the prior. This mix can be seen as a weighted average of both information. The weights depend on two elements, the strength of the prior (its precision) and the amount of data, with the weight of the prior decreasing as the number of subjects in the study increases. Obviously, the stronger the prior information, the more time it will take to lower its influence on the prediction (Gajewski & al., 2008; Jiang & al., 2016).

In this thesis, we decided to focus on the monitoring stage of a clinical trial when the prior information elicited by the trialists is combined with the observed data, to obtain an updated prediction of recruitment. As we mentioned, at the design stage of a trial, investigators have no choice but to provide information in the prior, otherwise, they cannot realise any prediction. However, as subjects are included in the trial, this prior information can have several impacts

on the prediction. Depending on the strength of the prior distribution, the posterior prediction will be weighted toward the prior (strong prior) or toward the observed data (weak prior). As an example, if no prior is introduced, the model will only rely on the data, whereas, if the researchers are highly confident on their prior, the model will require a large quantity of data to adapt prediction (Jiang & al., 2014). Hence, several questions arise regarding the prediction of recruitment in clinical trials:

- What would be the consequences of a strong informative prior which is significantly off target? It could lead to huge bias in the prediction of recruitment by sending reassuring signals to examiners. It would thus prevent them from detecting early warning signs of slow accrual, which, in return, alleviates their capacity to adapt and respond (Jiang & al., 2014). As a consequence, production planning for drugs would be inadequate, which would lead to increased costs for the firm.
- What are the benefits of a strong informative prior which is significantly on target? In case of a slow start in a trial, it could avoid unnecessary warning signs and prevent investigators from overreacting (Gajewski & al., 2008). It would give very accurate results, which would allow trialists to develop a robust planning as both sources of information would lead to the same results.
- For both questions, what are the differences if a weak prior is used instead? For the first question, the weak prior would allow the model to quickly adapt to the observed data and help investigators to make an appropriate corrective plan and more adequate drug distribution planning. However, in the context of the second question, the model would quickly react to early alarm and give misleading results.

This work aimed at addressing those questions at the start of the monitoring stage, when the quantity of data available is limited. The reason is that the methods using Bayesian approach to monitor an ongoing trial still face issues regarding the predictions at the beginning of a trial when few data are available (Gajewski & al, 2008; Zhang & Long, 2010; Jiang & al, 2014). Hence, in the context of drug distribution planning issues, the objective was to study the influence of the prior information on the ability to predict recruitment accurately as well as identifying what would be the most relevant quantity of information to provide. We wanted to confront the benefits and disadvantages of using a strong informative prior and compare it with a case where a weak prior was introduced.

Obviously, in a real clinical trial context, investigators have no idea if their prior is on target or not when designing the study. Their decision is thus based on their confidence and the valuable

knowledge they have about the prior choice. In this perspective, the relevant implications for trial teams concern the risks of using a strong prior at the start of a trial when it turns out to be off target. In other words, what is the "costs" of using a strong prior, which is off target, when a weak one could have been preferred?

3.2 Methodology

3.2.1 Assumptions

For the purposes of this thesis, we assumed that part of the information was fixed or known. The assumptions are the following:

- The sample size (n) and the total recruitment time ($T_{expected}$) were set, defined and endorsed by all the people in charge of the clinical trial.
- The data used were the waiting times (w_i, i=1,…,n) between the recruitment of subjects in the study. They correspond to the difference in days between the dates of arrival of two subsequent patients. Thus, $w_i = t_i - t_{i-1}$ for patient i and $t_0 = 0$ represents the start of the study.
- The waiting times are independent and identically distributed (i.i.d.).
- The recruitment occurred only in one center.
- The accrual rate was constant over time.

All those assumptions play an important role in the context of the analysis and the upcoming model presentation.

3.2.2 Model

The model we considered for the waiting times was first presented by Gajewski, Simon, and Carlson (2008). It aims at predicting the waiting times of the n subjects the trialists have planned to recruit (design stage) or to predict the remaining w_i, i=m+1,…,n, knowing that a number (m) of patients has already been recruited (monitoring stage). The waiting times are assumed to follow an exponential distribution with mean λ ($w_i \sim Exp(\lambda)$), implying, in accordance with the assumptions, a constant accrual rate. The density function of the w_i is:

$$f(w|\lambda) = \frac{1}{\lambda} e^{-\frac{w}{\lambda}} \tag{15}$$

3.2.2.1 Prior on λ

For the parameter λ, a conjugate prior for the exponential distribution is assumed: the inverse gamma distribution ($\lambda \sim IG(\alpha, \beta)$). The density function is the following:

$$f(\lambda|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{-\alpha-1}\exp\left(-\frac{\beta}{\lambda}\right) \quad (16)$$

For the purpose of the thesis, we needed to define values for α and β where the quantity of information introduced in the prior is easily identifiable. Two moments of the inverse gamma distribution seemed appropriate to achieve this, the mean [$\beta/(\alpha-1)$] and the variance [$\beta^2/((\alpha-1)^2(\alpha-2))$]. The mean sets the distribution around investigators expected value about λ while the variance shows the confidence on the values. To define those moments, two different methods were used:

- The mean: as λ represents the mean of the waiting times distribution, matching its expected value with the investigators' expected accrual rates in the study seems reasonable. The latter, which can also be referred as the expected waiting time, can be estimated by the total expected recruitment time (in month) divided by the sample size:

$$E(\lambda) = E(w_i) = \frac{\mathrm{T_{expected}} * (365.25/12)}{\mathrm{n}} \text{(in days)} \quad (17)$$

  As waiting times represent differences in days between consecutive patients' arrivals, it seems more consistent to present its expected value in the same unit. Thus, the total expected recruitment time is transformed in days, which explain the presence of the constant $\frac{365.25}{12}$.

- The variance: it represents the level of confidence investigators have on their prior. Contrary to the mean, it will not be a fixed quantity, and varies depending on the information put into the model. A low variance of λ in the model reflects a high degree of confidence in the prior among researchers. Conversely, the greater the variance, the more uncertainty is introduced into the a priori information. To illustrate this, several graphs are displayed in **Figure 3**. They show the distribution of λ with regard to different values of variance. As it can be seen, the larger the variance, the more spread λ possible values. Also to be noted, the range of variance values depends on the constant $T_{expected}$ and n.

Based on those two moment values, we obtain the following values for α and β:

$$\hat{\alpha} = \frac{E^2(\lambda)}{V(\lambda)} + 2 \quad (18)$$

$$\hat{\beta} = E(\lambda)\,(\hat{\alpha}-1) \quad (19)$$

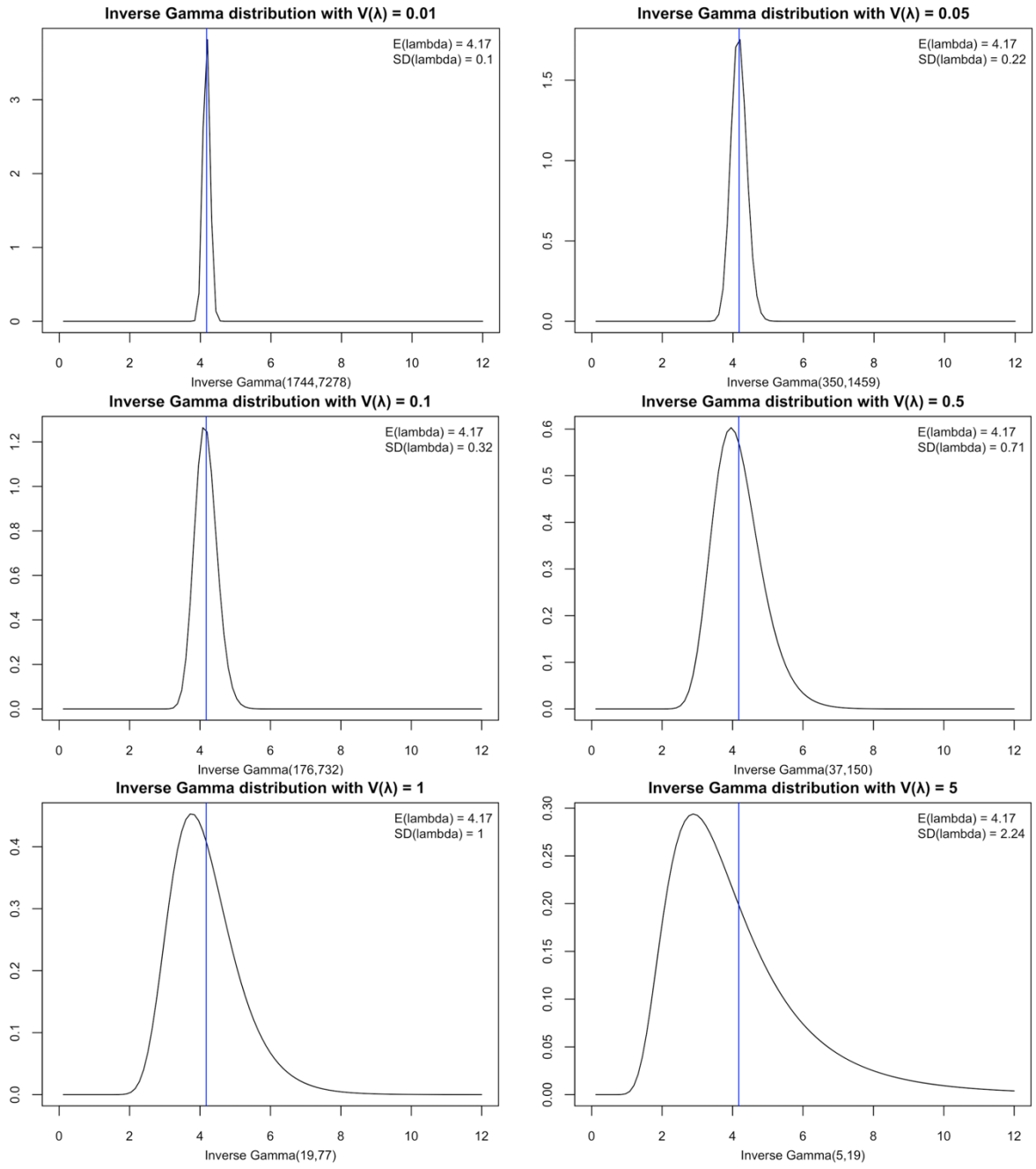

Figure 3. Influence of the variance on lambda prior density distribution. *The values used are n = 350, $T_{theoric}$ = 48 months and different V(λ) (0.01, 0.05, 0.1, 0.5, 1, 5). The expected waiting time, E(λ), equals 4.17 days.*

3.2.2.2 Posterior on λ

To obtain the posterior distribution of λ, we used Stan[3]. By means of MCMC sampling (see **Markov Chain Monte Carlo**), it provides an estimated posterior distribution based on the data

[3] Stan is an open-source program that performs Bayesian analysis with the use of MCMC approach. To be more precise, it generates samples from a parameter posterior distribution through Hamiltonian Monte Carlo simulation algorithm (Gelman & al., 2013).

from the m subjects already in the study. An analytical solution is also derivable, since we have a conjugate prior. It has the following form:

$$\lambda \sim IG(\hat{\alpha} + m, \hat{\beta} + \sum_{i=1}^{m} w_i) \quad (20)$$

Despite the fact that the MCMC method is more computationally and time demanding than the calculated analytical solution, we chose it for two reasons. (i) It is the method used by Pharmalex in their modeling. (ii) If we want to use more complex or different distributions for the prior or the likelihood, which involves no analytical solutions for the posterior, the method would already be set. It would only require small modifications of the code to change the distributions and obtain the desired results. This was done considering possible extensions of the model for future research.

3.2.2.3 Prediction algorithm

We were interested in predicting the number of subjects recruited per month in the trial. To accomplish this, the arrival dates of the remaining n-m patients needed to be drawn. We also had to take into account the uncertainty in the λ parameter and in the recruitment process. For the latter, the prediction process was repeated 1000 times, in order to obtain a number of arrivals by month distribution. The value of "1000" was chosen for one main reason: in the scientific literature dealing with accrual prediction of clinical trials, it represents the typical value used to derive the predictive distribution of a quantity of interest (Gajewski & al, 2008; Deng & al, 2014). For the uncertainty on the parameter, different values of λ were used to predict the waiting times. Two different prediction algorithms were used: (i) one for the situation before any patient has entered the trial (design stage) and (ii) one for the situation where some subjects have already been recruited (monitoring stage). Both algorithms are described below.

i. Design stage
    1. A vector of parameter values λ_i (i=1,…,n_pred) is randomly sampled from its prior distribution, $\lambda_i \sim Inv\ Gamma(\hat{\alpha}, \hat{\beta})$, where n_pred represents the number of subjects to be predicted.
    2. The waiting times, w_i (i=1,…,n_pred) are randomly drawn from the exponential model, with one λ_i generating one w_i, $w_i \sim Exp(\lambda_i)$.

3. Using the cumulative sum of the waiting times, the dates of randomisation are derived, which then allows to compute the number of subjects recruited per month.
4. Repeat the process 1000 times.

ii. Monitoring stage

1. A parameter values λ_t (t=1,…,1000) is selected from the posterior distributions sample generated in Stan. Here, 1000 corresponds to the size of the posterior distribution sample.
2. The waiting times w_i (i=1,…,n_pred) are randomly drawn from the exponential model, $w_i \sim Exp(\lambda_t)$.
3. Using the cumulative sum of the waiting times, the dates of randomisation are derived, which then allows to compute the number of subjects recruited per month.
4. Repeat the process 1000 times.

Eventually, it gives a complete distribution of the number of subjects recruited per month, with the possibility to compute credible intervals and probabilities. Moreover, it offers investigators the capacity to develop several recruitment scenarios each giving different arrival numbers.

3.2.3 Analysis and Results

The objectives were explored through the use of simulation studies. Each simulation corresponded to a clinical trial recruitment where the waiting times were generated from an exponential distribution. The parameter λ depended on the settings of the simulation, which involved two quantities: (i) n, the number of subjects to be recruited; and (ii) $T_{theoric}$, the theoretical number of months to recruit them. On this basis, the value of λ measured the theoretical mean waiting time, i.e. the theoretical number of days between two patients' arrival:

$$\lambda_{theoric} = \frac{T_{theoric} * (365.25/12)}{n} \text{(in days)} \quad (21)$$

Waiting times generation being random, it therefore provides a specific real recruitment duration, T_{real}, to obtain the n subjects in each simulation (in months). However, the values of T_{real} are on average close to $T_{theoric}$, as highlighted in **Table 3**. Practically, T_{real} is obtained by deriving the subjects' dates of randomisation using the cumulative sum of the waiting times

and then, by counting the number of months from the start of the trial to the last patient arrival. **Figure 4** illustrates an example of one simulation with n=350 and $T_{theoric}$ =48, meaning $w_i \sim Exp(\lambda_{theoric} = 4.17)$. In this case, the recruitment duration obtained was 50 months.

	T_{real} (in months)					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$T_{theoric}$ = 15	10	14	15	15.50	17	23
$T_{theoric}$ = 20	18	20	20	20.50	21	22
$T_{theoric}$ = 24	19	23	24	24.44	26	30
$T_{theoric}$ = 36	32	35	36.50	36.47	37	41
$T_{theoric}$ = 40	35	39	40.50	40.49	42	46
$T_{theoric}$ = 60	56	60	60	60.50	61	66

Table 3. Statistical summaries of the real recruitment duration distributions generated by the 1000 simulations with different values of $T_{theoric}$.

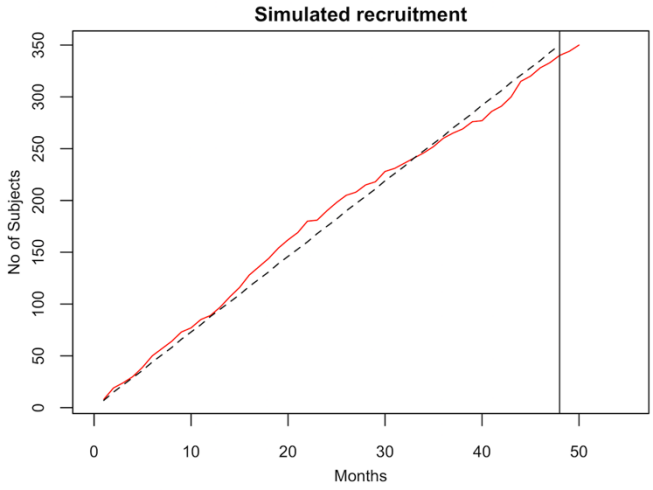

Figure 4. Accumulated accrual for a simulated recruitment. *The red line represents the cumulative sum of subjects recruited per month in the simulated data. The dashed line represents the theoretical accrual, if the recruitment was constant and corresponded to $n/T_{theoric}$ subjects per month. The vertical line represents the month in which the simulated recruitment should theoretically be completed (the 48^{th} month).*

The goal of these simulations was to explore situations where the examiners' prior beliefs on the recruitment duration, $T_{expected}$, was situated at various distances from T_{real} (in months, e.g. 0, 6 and 12 months), providing different prior values on the expected waiting times. The difference between $T_{expected}$ and T_{real} was referred as planned duration errors. The larger it is,

the more off target the prior is considered. Each one represented a different scenario. Then, in each scenario, we compared the results when different levels of prior information were introduced, going from the most informative to the least one, by changing the variance of the prior distribution of lambda. In the end, we wanted to assess which level of information was the most relevant given differences between T_{real} and $T_{expected}$. Note that, in this thesis, the planned duration errors from T_{real} involve overly optimistic recruitment durations, meaning that $T_{expected}$ is T_{real} minus the distance considered.

Based on the Bayesian knowledge, we had some intuitions about the scenarios outcomes:

- When $T_{expected}$ is far away from T_{real}, i.e. entirely off from the target, much better results are expected for the less informative priors. Informative priors require many subjects to enter the study to adjust to the true data.
- When $T_{expected}$ is close from T_{real}, i.e. on target, much better results are expected for more informative priors. Non-informative priors require more data to decrease the variability of estimates and become more accurate.

Thus, we wanted to confront the expected outcomes to the ones of the simulations to achieve several goals. Firstly, we wanted to corroborate those intuitions, because it might be, that the weak prior quickly performs as well as the strong one when $T_{expected}$ is close from T_{real}. In this case, it would make no sense to provide an informative prior. Secondly, if the first objective was to be confirmed, by seeking, at each scenario level, which quantity of information is the most appropriate, we wanted to understand when $T_{expected}$ can be considered close to the real recruitment time, referred as the "closeness". Overall, those two objectives combined aimed at (i) defining the most relevant level of information to be introduced into a model, and (ii) to observe if a more informative prior represents a good decision.

Four simulation studies were designed to mimic real clinical trials. Each one was repeated 1000 times and involved a unique setting, i.e. combination of n and $T_{theoric,}$ summarized in **Table 4**. In each setting, the goal was to provide robust outcomes to compare the performances of the levels of information introduced in the prior. As means of comparison, for each variance, we assessed the accuracy of the predictions at various time points (e.g. 3, 6, 9 and 12 months) as from the month where patient m was recruited. Then, to derive more general observations, we compared the results from one setting to another. It is important to note that we were not interested in knowing if the priors are accurate in predicting accrual, but rather in comparing

their accuracy when predicting accrual. We wanted to know which prior has the closest predictions to real recruitment in each scenario.

Before, describing the main characteristics of the settings, general comments needed to be made to explain the analysis.

- Two types of results were reported to assess the prediction accuracy of an information level at each time point:
    - The accuracy of the predictions based on the median number of subjects recruited per month compared to the true one (later referred as median precision). Here, we look at the mean square errors (MSE) between those two values over the 1000 simulations. At each time point for predictions, t, the MSE is:

$$MSE^t = \frac{1}{1000}\sum_{i=1}^{1000}(n_{i,real}^t - n_{i,median\ pred}^t)^2 \qquad (22)$$

      In addition to measuring the gap between real and predicted recruitment values, MSE takes into account the variability of the 1000 gaps computed and penalizes for huge differences. When comparing several priors, the one(s) with the smallest MSE will be considered as the one(s) with the best performance(s). We decided to work with the median instead of the mean or the mode, since we believe it represents a more robust measure of centrality. Moreover, it represents the measure generally used by Pharmalex in those situations.
    - The proportion of times the real recruitment value is included in the 90% credibility intervals. This result is referred as the correct coverage (CC) or coverage rate (CR). At each time point for predictions t, the correct coverage is:

$$CC^t = \frac{1}{1000}\sum_{i=1}^{1000} I(n_{i,lower\ bound}^t \leq n_{i,real}^t \leq n_{i,upper\ bound}^t) \qquad (23)$$

      When comparing several priors, the one(s) with the highest CC will be considered as the one(s) with the best performance(s). The credible intervals computed are the highest posterior density intervals, as presented and explained in **Credible Intervals (CI)**.

  Thus, we sought for variance with a small MSE in median prediction with a high correct coverage percentage.

- The scenarios always involved $T_{expected}$ set at 12, 6, 3 and 0 months from T_{real}. 12 and 0 were chosen as the two extremes to check whether the intuitions presented above were true, 12 being assumed as far away from T_{real} and 0 as close. Eventually, the last values, 6 and 3, were chosen to improve the definition of closeness. In other words, we wanted to narrow the interval of planned duration errors where an expected duration is considered close to the real duration. As an example, if weak priors were to perform the best in all situations considering a planned duration error of 6 months, we would know that closeness as to be situated between 0 and 5 months from the real duration. We chose to use 6 months as narrowing value based on a recent research on recruitment success (Pemberton & al., 2018). By analyzing the results on 167 trials, they found that 73.3% (123) were slower than expected. Moreover, of those 123, only 35% (43) were able to recruit more than 80% of their target sample sizes with an extension of 6 months. Thus, a difference of 6 months seemed like a reasonable value, since at least 47.9% of the trial made an error of 6 months or more in their planned duration. Eventually, 3 months was use to further narrow the interval.
- The values for m were selected to represent a small proportion of the data in the trial, at most 31% of the total sample size. They were chosen incrementally, to show the effect of the data on the different priors studied.
- The time points to assess prediction accuracy were situated at regular intervals (in months) from the month of the latest subject's arrival. They were chosen to evaluate the effect of priors as the predictions became more distant in time. The first points were more in the short-term, while other points were involving middle to long-term predictions. The periods considered for the time points were depending on the value of $T_{theoric}$.
- Regarding the level of information introduced in the prior, the values were chosen to present the gradual effect of information loss on the results. Moreover, we decided to separate them in two groups. They were either considered as a strong or a weak prior. The choice is made depending on the form of its density, as presented in **Figure 3**. When the shape is almost symmetric, close to the form of a normal distribution, the prior was referred as a strong one ($V(\lambda)=0.01, 0.05, 0.1$), whereas when the form start to broaden on the right, the prior was referred as a weak one ($V(\lambda)=0.5, 1, 5$). Note that this choice is completely subjective and was made to ease the presentation of results. However, we believed it would not affect the interpretations, since the last objective is to check what

levels of variance seem the most appropriate given a situation. If neither of the priors among strong and weak ones yield the best results, we thus know that the most appropriate variance would lie somewhere in between the two groups.

Below, we present the different characteristics of the settings (also summarized in **Table 4**). To set up the different values of n and $T_{theoric}$, we looked at the types of real trials used in the scientific literature in terms of durations and sample sizes. We also checked on decisions previously made by researchers when simulation studies had to be run (Gajewski & al., 2008; Zhang & Long, 2010; Mijoule & al., 2012; Jiang & al., 2014; Lan & al., 2019). The values that vary along with the setting were V(λ), m and the time points for computing the prediction accuracy:

Setting 1 The theoretical values were n=350 and $T_{theoric}$=48 ($\lambda_{theoric}$=4.17). For the levels of variance introduced the values were: V(λ)=0.01, 0.05, 0.1, 0.5, 1, 5. V(λ)=0.01, 0.05, 0.1 are referred as strong or more informative priors, while V(λ)=0.5, 1, 5 are referred as weak or less informative priors. The number of patients already in the study were 10, 20, 30, 40, 50, 100, corresponding to 3%, 6%, 9%, 11%, 14% and 29% of the total sample size. For the prediction accuracy, the time points were situated at 3, 6, 9, 12 and 15 months from the month where patient m was recruited.

Setting 2 The theoretical values were n=500 and $T_{theoric}$=20 ($\lambda_{theoric}$=1.22). For the levels of variance introduced the values were: V(λ)=0.0005, 0.001, 0.005, 0.01, 0.05, 0.1. V(λ)=0.0005, 0.001, 0.005 are referred as strong or more informative priors, while V(λ)=0.01, 0.05, 0.1 are referred as weak or less informative priors. The number of patients already in the study were 15, 30, 45, 60, 75, 150, corresponding to 3%, 6%, 9%, 12%, 15% and 30% of the total sample size. For the prediction accuracy, the time points were situated at 2, 4, 6, 8 and 10 months from the month where patient m was recruited.

Setting 3 The theoretical values were n=1800 and $T_{theoric}$=30 ($\lambda_{theoric}$=0.51). For the levels of variance introduced the values were: V(λ)=0.0005, 0.001, 0.005, 0.01, 0.05, 0.1. V(λ)=0.0005, 0.001 are referred as strong or more informative priors, while V(λ)=0.005, 0.01, 0.05, 0.1 are referred as weak or less informative priors. The number of patients already in the study were 50, 100, 150, 200, 250, 500, corresponding to 3%,

6%, 8%, 11%, 14% and 28% of the sample size. For the prediction accuracy, the time points were situated at 3, 6, 9, 12 and 15 months from the month where patient m was recruited.

Setting 4 The theoretical values were n=48 and $T_{theoric}$=24 ($\lambda_{theoric}$=15.22). For the levels of variance introduced the values were: V(λ)=0.1, 0.5, 1, 5, 10, 50. V(λ)=0.1, 0.5, 1 are referred as strong or more informative priors, while V(λ)=5, 10, 50 are referred as weak or less informative priors. The number of patients already in the study were 3, 4, 5, 6, 7, 15, corresponding to 6%, 8%, 10%, 12%, 15% and 31% of the sample size. For the prediction accuracy, the time points were situated at 1, 2, 3, 4 and 5 months from the month when patient m was recruited.

	Theoretical values	N° of simulation	Scenarios	Prior Variance (V(λ))	N° of subjects recruited (m)	Time points for predictions
Setting 1	n=350 $T_{theoric}$ =48 $\lambda_{theoric}$=4.17	1000 simulations	1. $T_{expected}$ at 12 months 2. $T_{expected}$ at 6 months 3. $T_{expected}$ at 3 month 4. $T_{expected}$ at 0 month	Strong: 0.01-0.05-0.1 Weak: 0.5-1-5	10-20-30-40-50-100	At 3, 6, 9 and 12 months from last arrival At $T_{expected}$
Setting 2	n=500 $T_{theoric}$=20 $\lambda_{theoric}$=1.22	1000 simulations	1. $T_{expected}$ at 12 months 2. $T_{expected}$ at 6 months 3. $T_{expected}$ at 3 month 4. $T_{expected}$ at 0 month	Strong: 0.0005-0.001-0.005 Weak: 0.01-0.05-0.1	15-30-45-60-75-150	At 2, 4, 6, 8 and 10 months from last arrival
Setting 3	n=1800 $T_{theoric}$=30 $\lambda_{theoric}$=0.51	1000 simulations	1. $T_{expected}$ at 12 months 2. $T_{expected}$ at 6 months 3. $T_{expected}$ at 3 month 4. $T_{expected}$ at 0 month	Strong: 0.0005-0.001 Weak: 0.005-0.01-0.05-0.1	50-100-150-200-250-500	At 3, 6, 9, 12 and 15 months from last arrival
Setting 4	n=48 $T_{theoric}$=24 $\lambda_{theoric}$=15.22	1000 simulations	1. $T_{expected}$ at 12 months 2. $T_{expected}$ at 6 months 3. $T_{expected}$ at 3 month 4. $T_{expected}$ at 0 month	Strong: 0.1-0.5-1 Weak: 5-10-50	3-4-5-6-7-15	At 1, 2, 3, 4 and 5 months from last arrival

Table 4. Summary of the different simulations studies*. For each one, there are always three scenarios, six variance values, six m values and five time points for predictions. As a reminder, a scenario represents a specific difference between the examiners' prior beliefs on the recruitment duration, $T_{expected}$, and the real one T_{real}. $T_{expected}$ is T_{real} minus the difference considered.*

3.3 Simulation Studies

For this part, we will first present and describe in detail the results of each scenario obtained in the first setting. Then, for the remaining settings, we will just highlight the differences that are observed. Eventually, we will make general comments on the results.

To present the different results, graphics were selected as they represent a more convenient method than rough tables (available in **Annexe 2**). One graph per m value is displayed to show the performances of each prior, allowing direct comparisons. Note that the graph for the prior m=0 is included for information and will not be analysed, since we are studying adaptation of priors after the start of subjects' recruitment. Moreover, for the settings 2, 3 and 4, we only presented and displayed the results that differ from the ones already observed, e.g. if the setting 3 has similar results to setting 2, we will not elaborate on it. The graphics that were not shown are also available in **Annexe 2**.

3.3.1 First setting

For this setting, the theoretical recruitment duration was 48 months and the number of patients to be recruited was 350. **Table 5** provides the statistical summary of the real recruitment durations generated by the 1000 simulations.

T_{real} (in months)					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
41	47	49	48.56	50	57

Table 5. Statistical summary of the real recruitment duration distribution generated by the 1000 simulations simulations in the first setting.

3.3.1.1 Scenario 1: $T_{expected}$ at 12 months

For the first scenario, $T_{expected}$ was always set at 12 months from the real durations generated to recruit the n subjects. We considered these $T_{expected}$ as far away from the real recruitment durations, and, thus, expected the less informative prior to perform better. The first result to be reported is the MSE displayed in **Figure 5**. We were looking for priors with the smallest MSE values. The following observations can be made:

- At m=10, for the predictions at 3 months, there is not much difference among MSE of all priors. However, the further the predictions, the larger the differences, specifically between V(λ)=5 and the other values. We also notice that V(λ)=0.5 obtain the lowest MSE.
- At m=20, for the predictions at 3 months, again, MSE are very close. Nevertheless, the further in time the predictions are made, the greater the gap between the priors, with a clear advantage towards the less informative priors, V(λ)=0.5 and V(λ)=1. Moreover, V(λ)=5 obtain much better MSE values than at m=10. Indeed, its curve is now just over the ones of V(λ)=0.5 and V(λ)=1 and under the ones of strong priors.
- At m=30, 40, 50 and 100, the trends observed at m=20 are confirmed, with the difference that V(λ)=5 yields the same MSE curve as V(λ)=0.5 and V(λ)=1.
- Regarding more informative priors, the stronger it is, the larger the MSE at each m value.

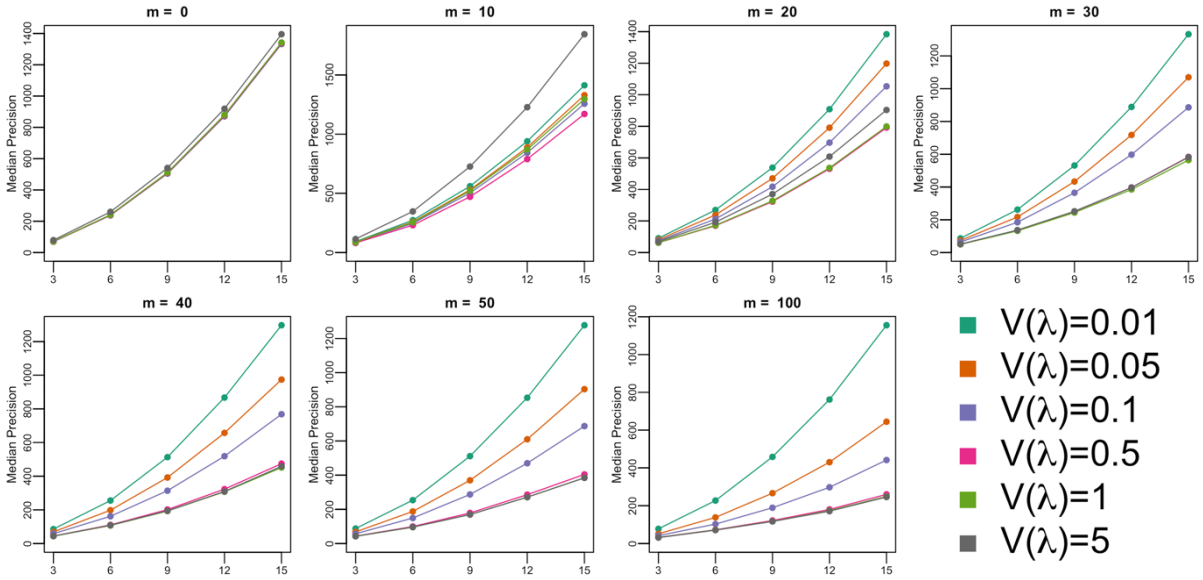

Figure 5. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the MSE for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at* $T_{expected}$*. m=0 corresponds to the prior. Note that, at m=30, 40, 50 and 100, results of V(λ)=5 completely or partially overlap either V(λ)=0.5 and V(λ)=1 or both.*

The second results to be reported, the correct coverage is displayed in **Figure 6**. We were looking for priors with the highest coverage rates. In this scenario, results look similar for all m values, with a clear advantage for less informative priors, specifically for V(λ)=1 and V(λ)=5 when m is small (10, 20 and 30). However, at m=40 50 and 100, V(λ)=0.5 obtains a closer

coverage to V(λ)=1 and V(λ)=5. Regarding the more informative priors, we observe that, the stronger it is, the lower its coverage rates.

Another interesting pattern to note in **Figure 6** concerns the evolution of the CC. As the number of subjects in the study increase, the curves of the strong priors tend to go up, whereas the ones of the weak priors tend to remain steady. We believe this can be explained by two factors, the evolution of prediction errors and credible interval widths:

- In the case of strong priors, the CI width evolution over m is slow, due to the little variations introduced by the prior in the beginning. Thus, their coverage rates are better at each m value, since the errors in predictions decrease while the CI remain relatively constant.
- In contrast, in the case of weak priors, CI widths decrease more significantly over m. Indeed, the high variations introduced by the prior in the beginning are more impacted by data addition than strong priors, which reduce the uncertainty around predictions. Therefore, their coverage rates remain relatively steady, as the reduction of errors in predictions is balanced with the decrease of CI widths. In other words, for small m values, larger errors are covered by a larger interval, whereas it is not more the case when the number of subjects in the study increase. This is compensated by a diminution of errors in predictions over m.

All those observations taken into account, one exception remains in the case where V(λ)=0.5. It represents the unique prior which has a high coverage rates already at m=10 and keeps on improving as m increases until it catches up the coverage of V(λ)=1 and V(λ)=5. This could be explained by the balancing between CI widths and prediction errors, where predictions become more accurate, while still having a high variation.

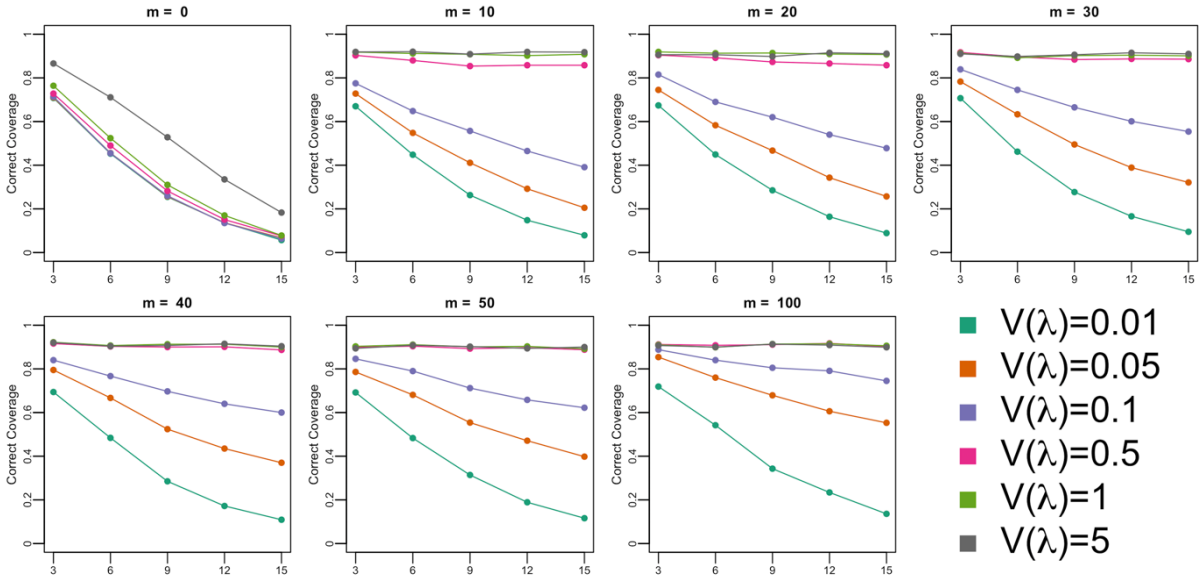

Figure 6. The number of times the real recruitment value is included in the 90% credibility intervals. *One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the coverage rates for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at m=30, 40, 50 and 100, results of V(λ)=5 completely or partially overlap either V(λ)=0.5 and V(λ)=1 or both.*

3.3.1.2 Scenario 2: $T_{expected}$ at 6 months

For the second scenario, $T_{expected}$ was always set at 6 months from the real durations generated to recruit the n subjects. The MSE is displayed in **Figure 7**. The following comments can be made:

- At m=10, 20 and 30, for the predictions at 3 months, there is not much difference among MSE of all priors, except V(λ)=5. However, the further the predictions, the larger the differences, specifically between V(λ)=5 and the other values. Regarding the performances, strong priors obtain the lowest MSE.
- At m=40 and 50, the MSE of weak priors get closer to the strong priors. More precisely, at m=40, V(λ)=0.5 obtains relatively the same MSE as V(λ)=0.01 and then a lower one at m=50. Strong priors still have the best MSE overall, specifically, V(λ)=0.05 and V(λ)=0.1.
- At m=100, for the predictions at 3 months, there is not much difference among MSE of all priors. However, for remaining predictions, there is split between V(λ)=0.01 and the other values. Moreover, V(λ)=0.5 and V(λ)=1 gets closer to V(λ)=0.1 and better than V(λ)=0.05, while V(λ)=5 has relatively the same MSE as V(λ)=0.05. Thus, in this scenario, we clearly observe a shift in the performances of priors when the amount of

data increases. The MSE of V(λ)=0.01 represents the best illustration, being among the lowest at m=10 and 20, then increasing slightly over the m values, to finish as the worst at m=100.

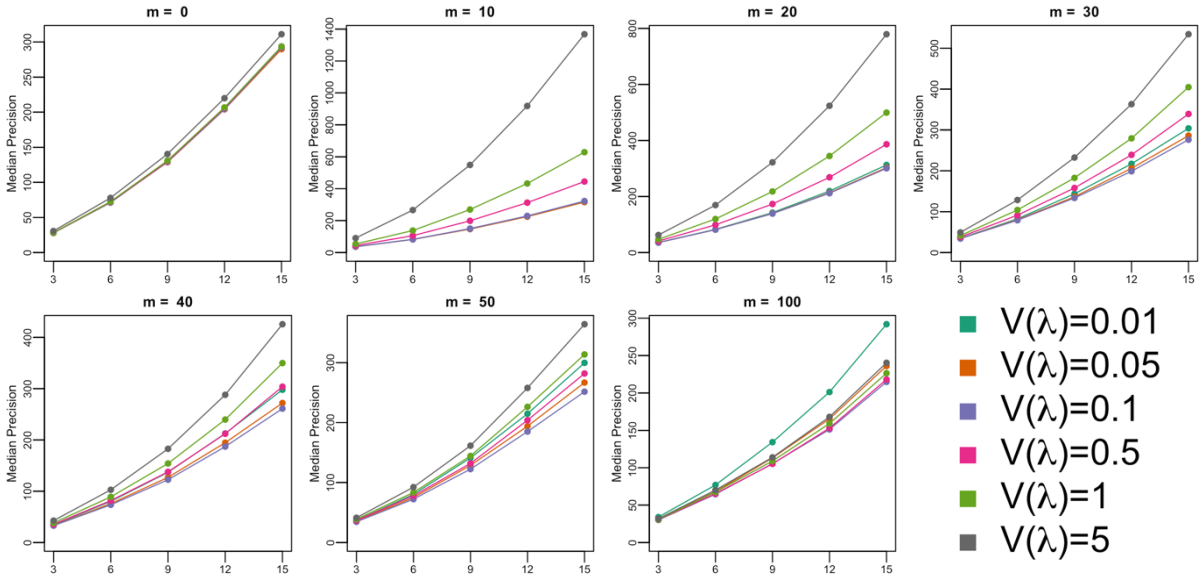

Figure 7. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the MSE for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at m=10 and 20, results of V(λ)=0.1 completely or partially overlap either V(λ)=0.05 and V(λ)=0.01 or both.*

The reported CC is displayed in **Figure 8**. In this scenario, V(λ)=0.5 and V(λ)=1 obtain the best coverage rates over all m values. Moreover, regarding the more informative priors, we observe that the stronger it is, the lower its coverage rates. However, some differences with scenario one are noted:

- The less informative prior, V(λ)=5, is no more among the best CC, but just below.
- The CC of strong priors are better, specifically when the predictions are close in time like 3, 6 and 9 months (CC over 80%). For other predictions, a slight drop for the CC was observed (below 80%). The only exception is V(λ)=0.01, which coverage rates do not decrease as predictions advance in time.
- V(λ)=0.01 obtains CC values very close to the ones of weak priors at all m values.

Concerning the evolution of the CC, like in the first scenario, the curves of the strong priors tend to go up with the increase of subjects in the study, whereas the ones of the weak priors tend to remain steady. In this scenario, V(λ)=0.5 remains relatively constant over m values, obtaining the best coverage performances for all m values.

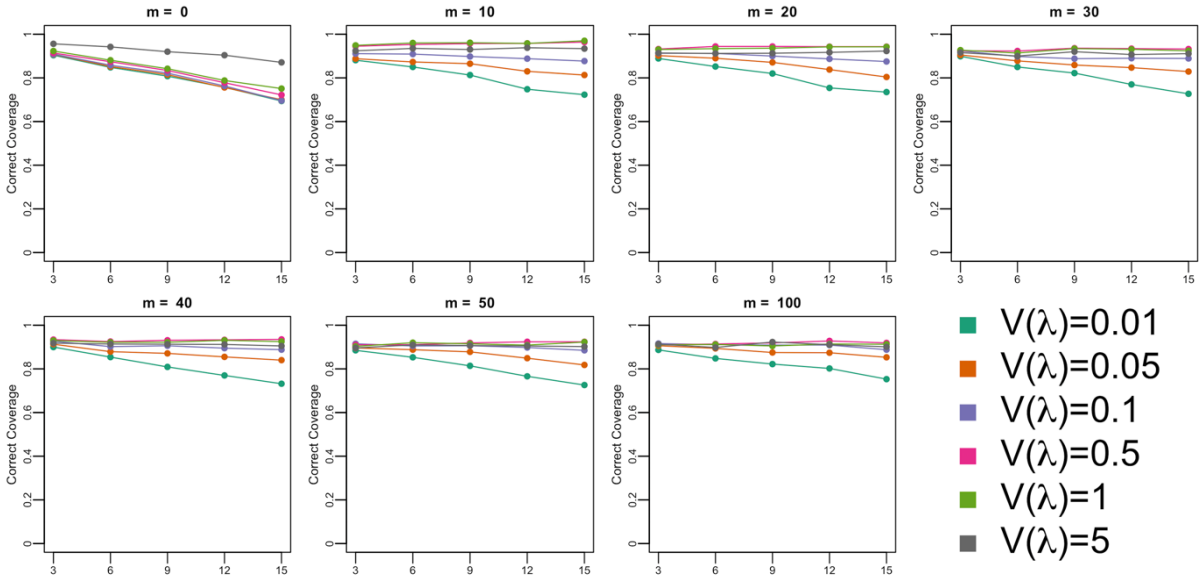

Figure 8. The number of times the real recruitment value is included in the 90% credibility intervals. *One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the coverage rates for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at m=10, 20 and 30, results of V(λ)=1 completely or partially overlap V(λ)=0.5. Also, at m=50 and 100, the results of V(λ)=5 completely or partially overlap V(λ)=0.1, V(λ)=0.5 and V(λ)=1.*

3.3.1.3 Scenario 3: $T_{expected}$ at 3 months

For the third scenario, $T_{expected}$ was always set at 3 months from the real durations generated to recruit the n subjects. The MSE are shown in **Figure 9**. Overall, all the graphics present the same kind of results. Except for V(λ)=5, the MSE are very similar in the 3-month predictions, then as the predictions become more distant in time, weak priors MSE increases more rapidly than the ones of strong priors. However, as m increase, the MSE of weak priors tend to get closer to the ones of strong priors.

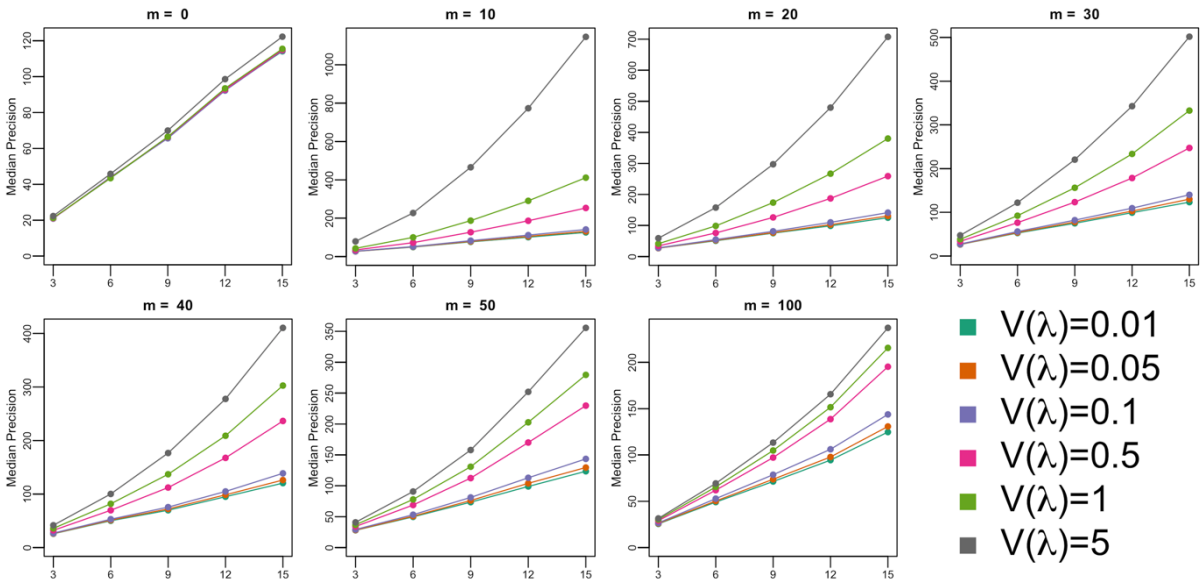

Figure 9. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the MSE for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at all m values, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05 and V(λ)=0.01.*

Concerning the coverage rates, displayed in **Figure 12**, on the whole, all priors obtain similar results, with a slight advantage towards V(λ)=0.5 and V(λ)=1. If we look at CC evolution, as m increase, the coverage rates of weak priors tend to decrease slightly, while the one of strong prior remain steady. This is again connected with the balance between CI widths evolution and prediction errors as explained above. For weak priors, the diminution of predictions errors is not big enough to offset the reduction of CI widths. For strong priors, CI widths remains constant over m values as well as the predictions errors, thus the coverage does not change much.

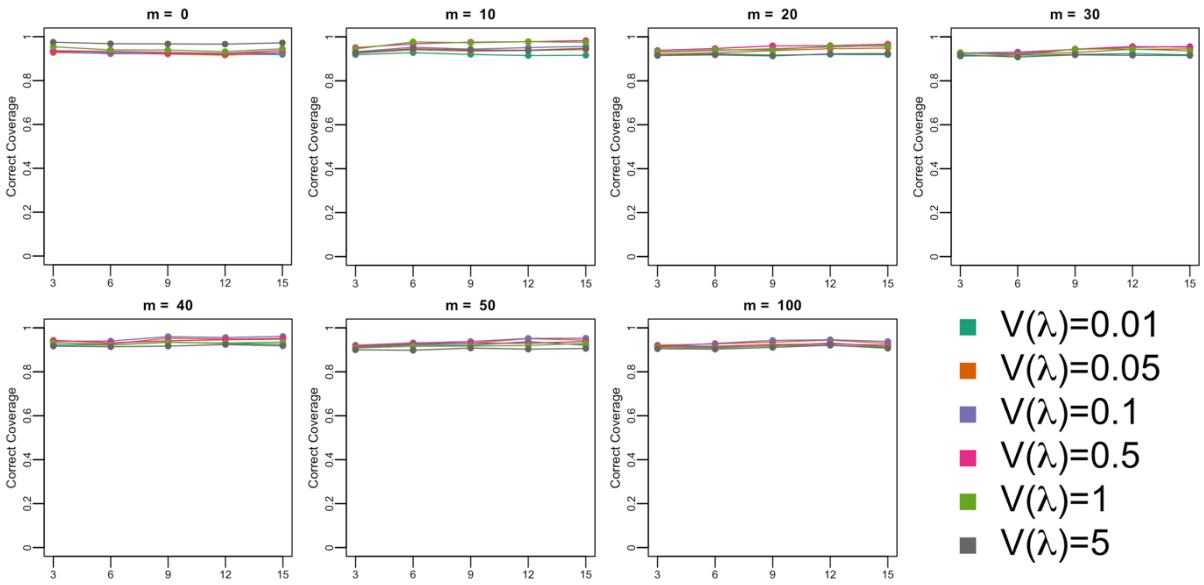

Figure 10. The number of times the real recruitment value is included in the 90% credibility intervals*. One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the coverage rates for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at* $T_{expected}$*. m=0 corresponds to the prior.*

3.3.1.4 Scenario 4: $T_{expected}$ at 0 months

For the fourth scenario, there was no difference between $T_{expected}$ and real durations generated to recruit the n subjects. Hence, we would expect more informative prior to perform better. The MSE are shown in **Figure 11**. Overall, all the graphics present the same kind of results as in the third scenario. Thus, as expected, more informative priors obtain the lowest MSE values. Similarly, the observations for the coverage rates, displayed in **Figure 12**, are the same as in the third scenario. The only difference is that, in this case, it is the strong priors that have a slightly better coverage than weak priors.

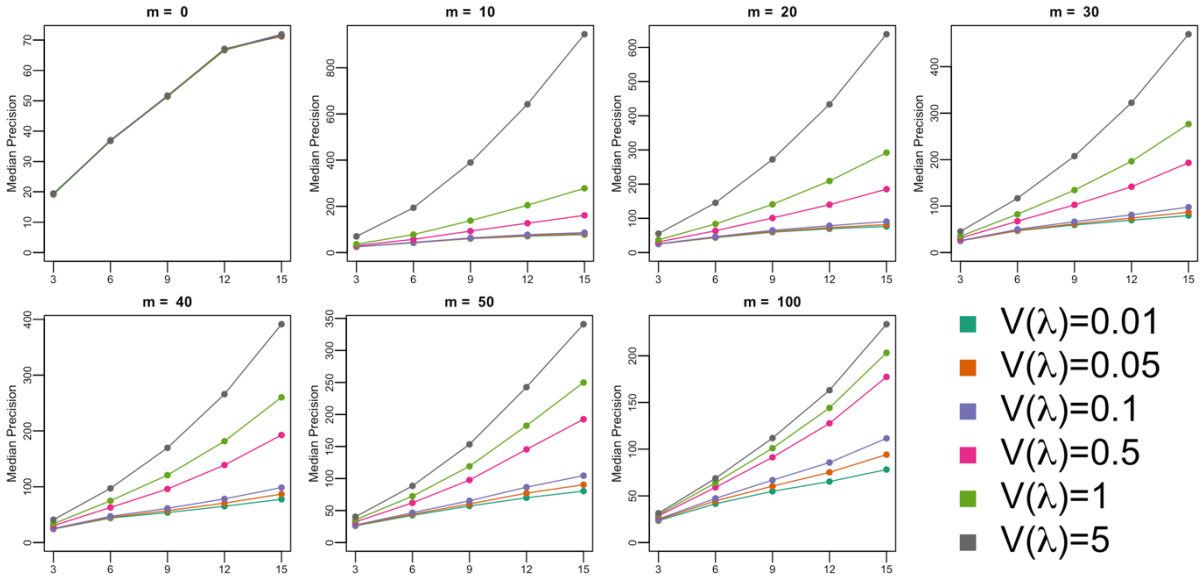

Figure 11. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the MSE for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at all m values, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05 and V(λ)=0.01.*

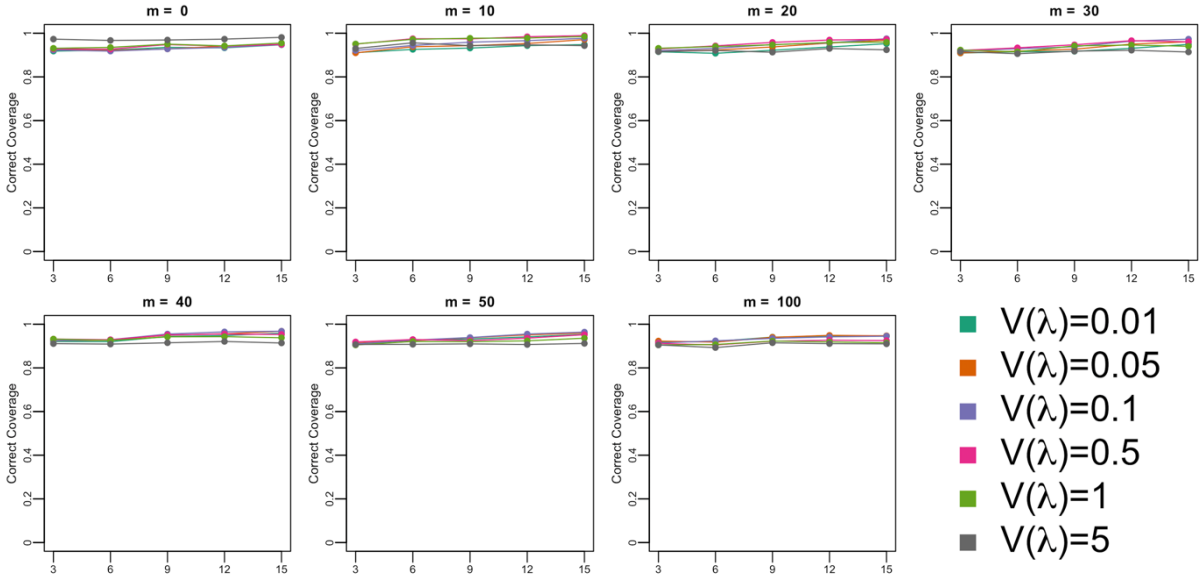

Figure 12. The number of times the real recruitment value is included in the 90% credibility intervals*. One graph is displayed per m value (0, 10, 20, 30, 40, 50, 100). It shows the coverage rates for the different priors at 3, 6, 9 and 12 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior.*

3.3.2 Second setting

For this setting, the theoretical recruitment duration was 20 months and the number of patients to be recruited was 500 ($\lambda_{theoric}$=1.22). **Table 6** provides the statistical summary of the real

recruitment durations generated by the 1000 simulations. The first, second and third scenarios provided different results from the first setting, while the results and observations of the fourth scenario were similar and, thus, not presented (available in **Annexe 2.2)**.

T_{real} (in months)					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18	20	20	20.50	21	23

Table 6. Statistical summary of the real recruitment duration distribution generated by the 1000 simulations in the second setting.

For the first scenario, the results have changed compared to the first setting. Regarding the MSE shown in **Figure 13**, the weak priors obtain the lowest values over all m. For the coverage displayed in **Figure 14**, the only difference concerns the case of the "stronger" variance in the weak prior group, V(λ)=0.01, since its CC is no more as high as it was in the first setting. Indeed, it was always around 90% in setting 1, while never being over 80% in this one, except at m=150. Eventually, looking at the evolution of the coverage with m, it corresponds to the observations of the first setting, as it was either increasing or remaining constant.

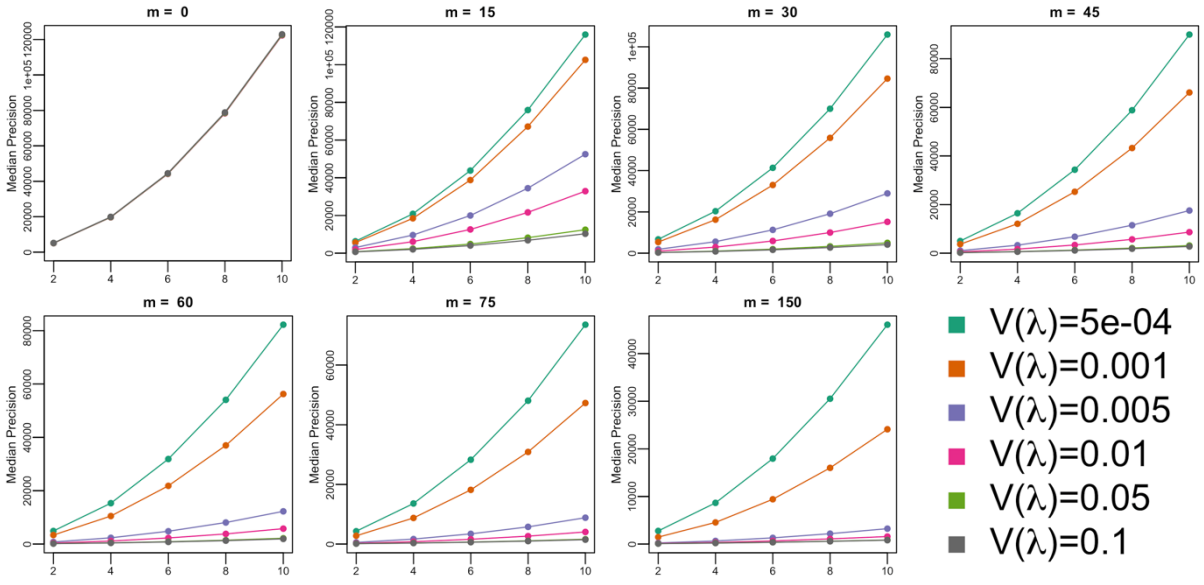

Figure 13. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 15, 30, 45, 60, 75, 150). It shows the MSE for the different priors at 2, 4, 6, 8 and 10 months from the month where patient m was recruited. m=0 corresponds to the prior. Note that, at all m values, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05. At m=75 and 150, results of V(λ)=0.1 partially overlap V(λ)=0.01.*

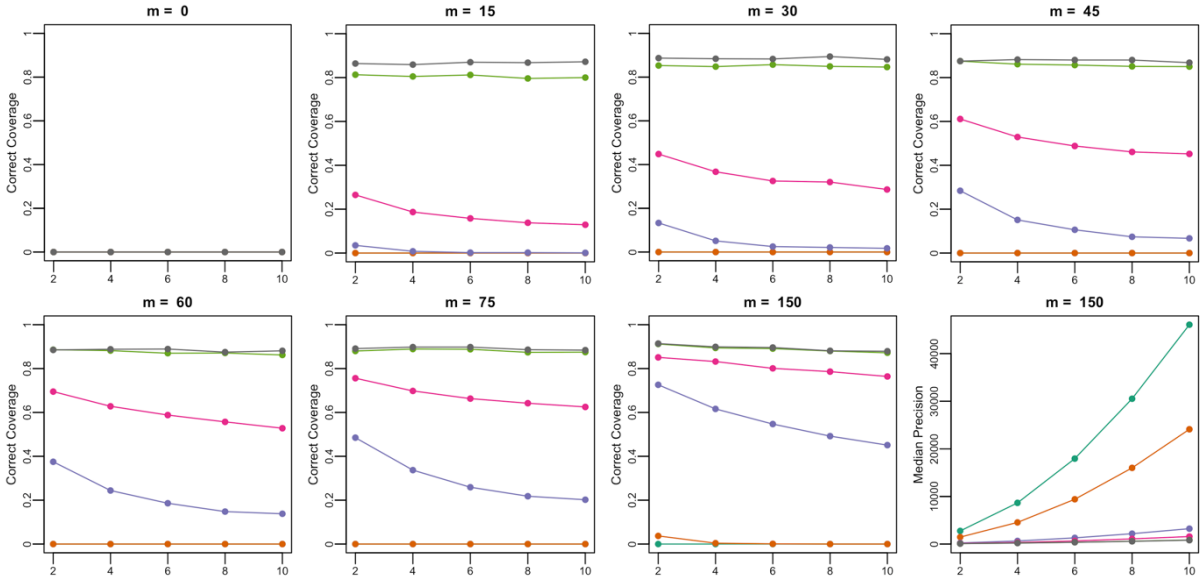

Figure 14. The number of times the real recruitment value is included in the 90% credibility intervals. *One graph is displayed per m value (0, 15, 30, 45, 60, 75, 150). It shows the coverage rates for the different priors at 2, 4, 6, 8 and 10 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at m=60, 75 and 100, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05.*

For the second scenario, a significant difference appeared between the results of the two settings. Unlike the first setting where MSE was the lowest for strong priors for small m values, in this setting, the less informative priors obtain the lowest MSE at each m value. This is illustrated in **Figure 15**. Moreover, regarding the coverage displayed in **Figure 16**, the differences between weak and strong priors are much larger than it was in the first setting over all m. The strong priors hardly reach 60% of coverage rates, except V(λ)=0.005 at m=150, while weak priors are always around 90%. The only exception concerns the "stronger" prior in the weak group, as its coverage rates do not match the ones of the weakest priors, V(λ)=0.05 and V(λ)=0.1.

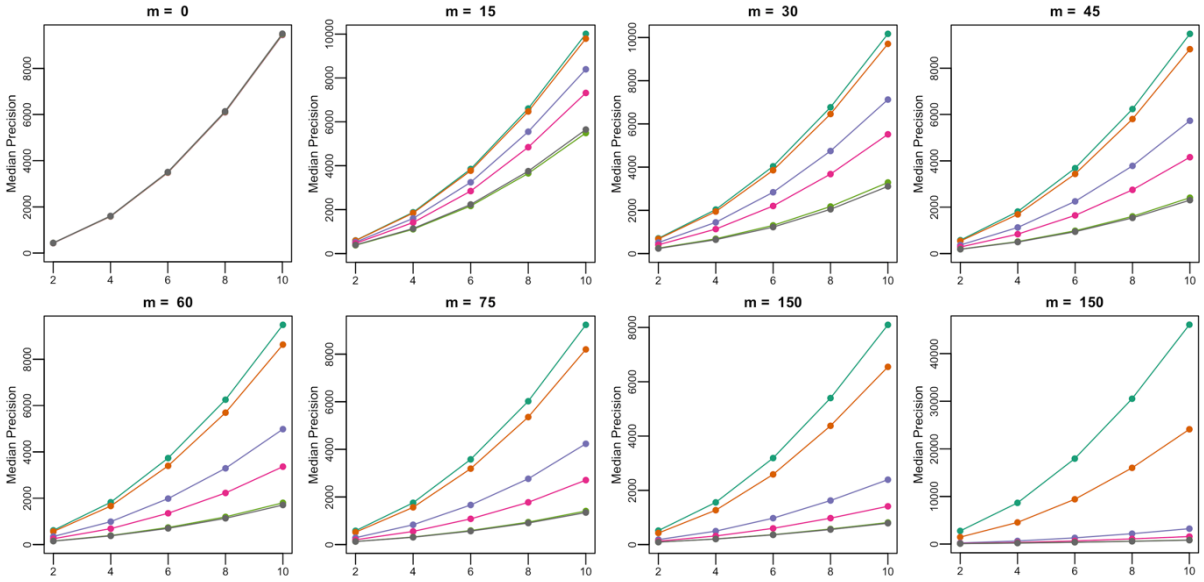

Figure 15. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 15, 30, 45, 60, 75, 150). It shows the MSE for the different priors at 2, 4, 6, 8 and 10 months from the month where patient m was recruited. m=0 corresponds to the prior. Note that, at all m values, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05.*

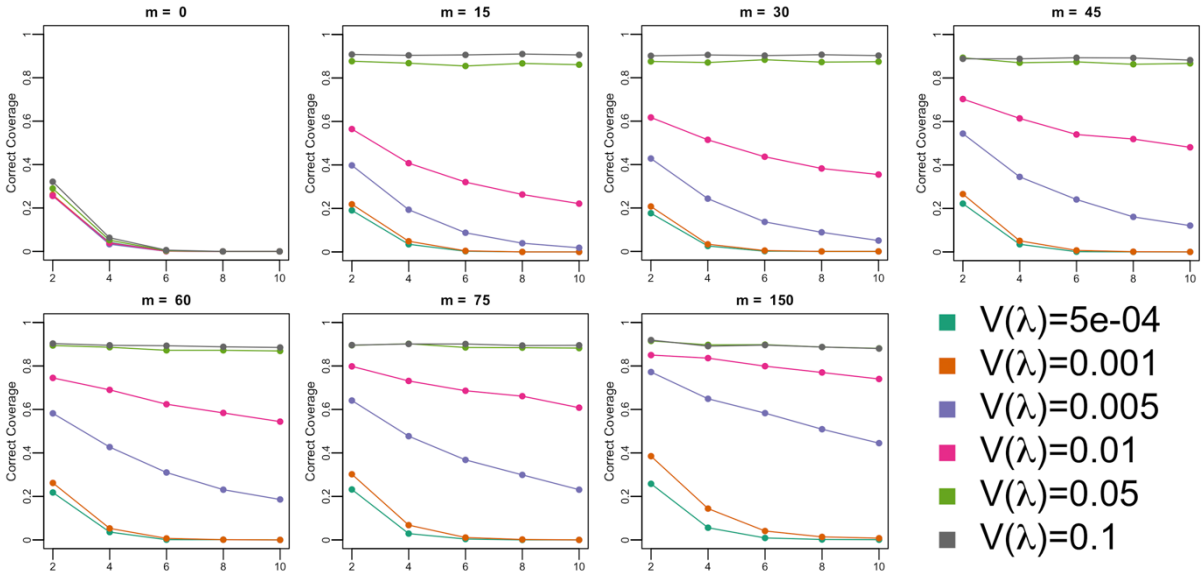

Figure 16. The number of times the real recruitment value is included in the 90% credibility intervals*. One graph is displayed per m value (0, 15, 30, 45, 60, 75, 150). It shows the coverage rates for the different priors at 2, 4, 6, 8 and 10 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at m=60, 75 and 100, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05.*

For the third scenario, we also observe a significant change. Regarding the MSE in **Figure 17**, results are similar to the second scenario of the first setting with strong priors that obtain the lowest MSE for small m values, then a shift in performance appears as m increases. Nevertheless, regarding the coverage rates in **Figure 18**, similarly to the second scenario of this

setting, the strong priors are obtaining low values (barely over 80% except for V(λ)=0.005) compared to the two weakest priors.

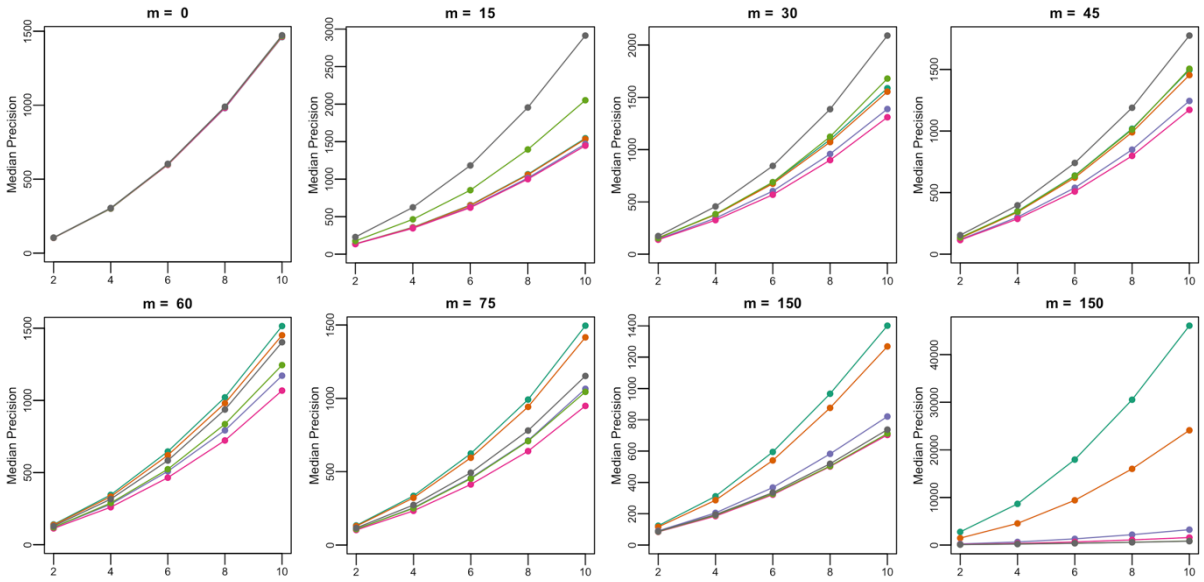

Figure 17. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 15, 30, 45, 60, 75, 150). It shows the MSE for the different priors at 2, 4, 6, 8 and 10 months from the month where patient m was recruited. m=0 corresponds to the prior. Note that, at m=15, V(λ)=0.01 completely or partially overlap V(λ)=0.005 and the two lower variances. The same at m=30 and 45, for V(λ)=0.05 on V(λ)=0.001 and V(λ)=0.0005; at m=75, for V(λ)=0.05 on V(λ)=0.005; at m=100 for V(λ)=0.1 on V(λ)=0.01 and V(λ)=0.05.*

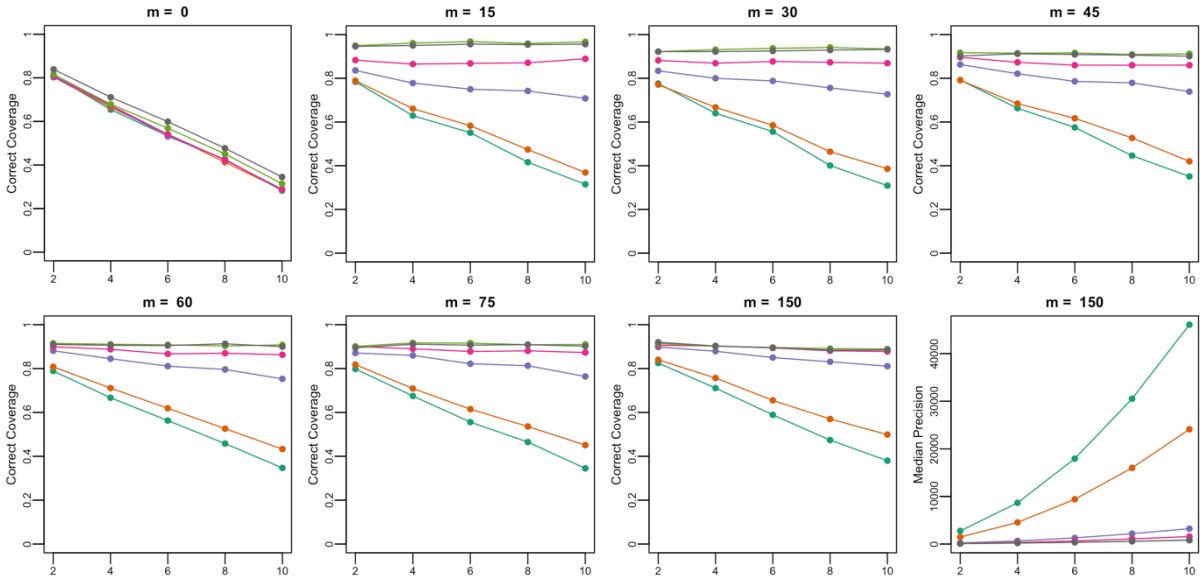

Figure 18. The number of times the real recruitment value is included in the 90% credibility intervals*. One graph is displayed per m value (0, 15, 30, 45, 60, 75, 150). It shows the coverage rates for the different priors at 2, 4, 6, 8 and 10 months from the month where patient m was recruited and at $T_{expected}$. m=0 corresponds to the prior. Note that, at m=60, 75 and 100, results of V(λ)=0.1 completely or partially overlap V(λ)=0.05.*

3.3.3 Third setting

For this setting, the theoretical recruitment duration was 30 months and the number of patients to be recruited was 1800 ($\lambda_{theoric}$=0.51). **Table 7** provides the statistical summary of the real recruitment durations generated by the 1000 simulations. The results and observations are the same as the ones of the second setting in the first, second and fourth scenarios, namely, better MSE and coverage rates for the weak priors in the first two scenarios, while strong priors performed better in the last scenario. However, for the third scenario, the results are highly similar to the second scenario of the first setting with strong priors having the best MSE and coverage rates close to the ones of weak priors. All the graphics are available in **Annexe 2.3**.

T_{real} (in months)					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28	30	31	30.53	31	33

Table 7. Statistical summary of the real recruitment duration distribution generated by the 1000 simulations in the third setting.

3.3.4 Fourth setting

For this setting, the theoretical recruitment duration was 24 months and the number of patients to be recruited was 48 ($\lambda_{theoric}$=15.22). **Table 8** provides the statistical summary of the real recruitment durations generated by the 1000 simulations. Only the results of the second scenario are different from the ones observed above. For the first, third and the fourth scenarios, the results and observations correspond to the ones of the first setting (**Annexe 2.4**).

T_{real} (in months)					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16	22	24	24.60	27	39

Table 8. Statistical summary of the real recruitment duration distribution generated by the 1000 simulations in the fourth setting.

For the second scenario, the results, shown in **Figure 19**, behave as in the first setting for the median precision. The strong priors obtain the lowest MSE at all m values except at the last one (m=15). In the first setting, the same was observed with weak priors starting to perform better

than strong ones from m=50 to m=100, the last m value. Nonetheless, in this case, the correct coverage is different. Indeed, as displayed in **Figure 20**, the CC of strong priors stay very close to the ones of weak prior over all m values, more than they were in the first setting. Just a slight drop in coverage is observed for strong priors as predictions are further in time, but they still stay over 85%.

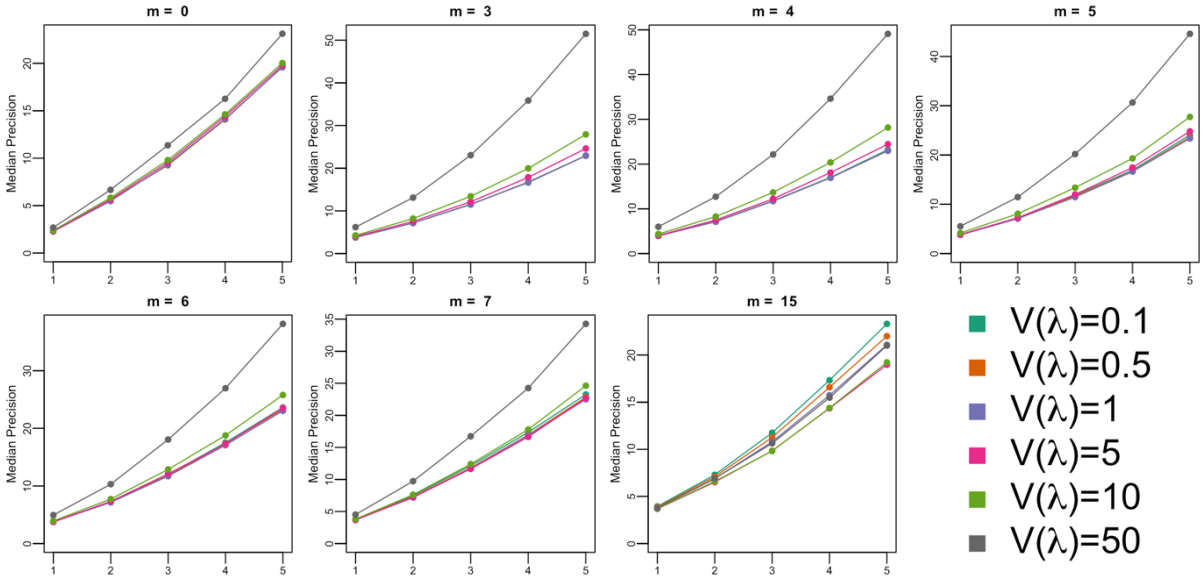

Figure 19. Prediction accuracy based on the median number of subjects recruited per month. *One graph is displayed per m value (0, 3, 4, 5, 6, 7, 15). It shows the MSE for the different priors at 1, 2, 3, 4 and 5 months from the month where patient m was recruited. m=0 corresponds to the prior. Note that, at all m values, results of V(λ)=1 completely or partially overlap V(λ)=0.1 and V(λ)=0.5. Moreover, at m=6 and 7, V(λ)=5 partially overlap V(λ)=1 and, at m=15, V(λ)=10 partially overlap V(λ)=5 and V(λ)=50 partially overlap V(λ)=1.*

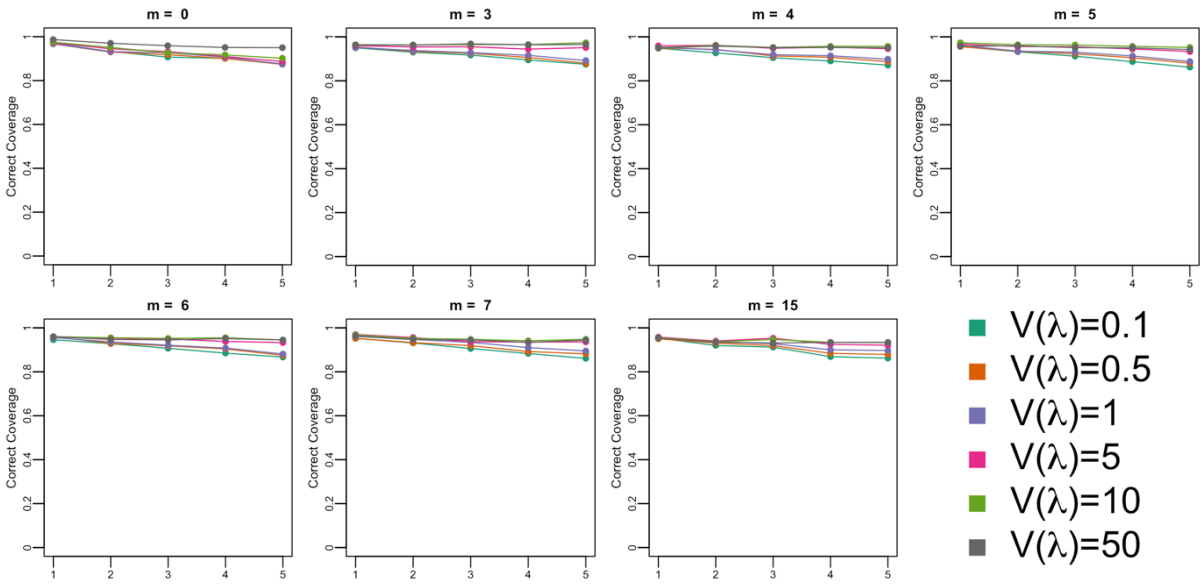

Figure 20. The number of times the real recruitment value is included in the 90% credibility intervals*. One graph is displayed per m value (0, 3, 4, 5, 6, 7, 15). It shows the coverage rates for the different priors at 1, 2, 3, 4 and 5 months from the month where patient m was recruited and at* $T_{expected}$*. m=0 corresponds to the prior.*

3.3.5 Global observations on simulation studies

To conclude with the simulation studies, we will link the results with the objectives. In doing so, we will sum up the observations from each setting. The first goal was to corroborate some intuitions about scenarios outcomes. As a reminder, we were expecting the following:

- When $T_{expected}$ is far away from T_{real}, uninformative priors should show the most accurate predictions.
- When $T_{expected}$ is close to T_{real}, much better results should be observed for the strong priors, in terms of predictions accuracy.

Based on the simulation studies, we could confirm those intuitions. In all the first scenarios, when $T_{expected}$ was assumed far away from T_{real}, the weak priors obtained overall better accuracy results than more informative priors. They had the lowest MSE and, at all m values, the highest coverage rates. Similarly, in the fourth scenarios where $T_{expected}$ was considered close to T_{real}, the opposite was observed, with a better performance from more informative priors.

The second goal was to understand when $T_{expected}$ could be considered as close to T_{real}, by seeking, at each scenario level, which quantity of information was the most appropriate. The first and fourth scenarios of each setting provided relevant information concerning this objective, defining the first boundaries. We knew that a close $T_{expected}$ would be located between 0 and 11 months. To complete the investigation, we used the second and third scenarios. It

reached different outcomes depending on the setting. To be more precise, we were confronted with three types of situations:

- In the first setting, for the third scenario, it was clear that the strong priors had the best results. However, for the second scenario, it was more complicated to settle on whether weak priors or strong priors obtained the best overall performance. Indeed, the priors that had the best balance of results if we look at MSE and Coverage over m values were the two central variances, $V(\lambda)=0.1$ or $V(\lambda)=0.5$. The advantage of more informative priors was clearly a better accuracy in terms of MSE, but they obtained poorer coverage than less informative priors.
- In the second setting, for both scenario, less informative priors clearly obtained the highest coverage over all m values. Moreover, they had the lowest MSE on all (second scenario) or on a part (third scenario) of m values.
- In the third setting, for the second scenario, less informative priors clearly obtained the best results. However, in the third scenario, results were mitigated as in the second scenario of the first setting. We could not decide which group of priors was the most relevant.
- In the fourth setting, more informative priors had the best performances, with lower MSE than weak priors and similar coverage at all m values in both scenarios.

Therefore, we faced contradictory information, as the closeness was not the same across settings. Based on our assumptions, in the first setting the interval was between 0 and 6 months. In the second setting setting, it was between 0 and 2 months. In the third setting, it was between 3 and 5 months. Eventually, in the fourth setting, it was between 7 and 11 months.

The last objective was subdivided in two parts. The first part consisted in identifying the most relevant levels of information to be introduced into a model. To explore it, we looked at which priors gave the best results over all scenarios in each setting. This thesis being more focused on drug availability issues in clinical trials, we decided to favor priors that ensure high coverage in all scenarios. Then, to discriminate among them, we selected the ones having or being close to the best MSE across scenarios. Based on the results, here are the priors identified:

Setting 1 $V(\lambda)=0.5$, as it ensured a high coverage in all scenarios, while being among the best MSE in scenario one. Moreover, for the second scenario, MSE is really close to the ones of strong priors for m=10, 20, 30 and 40, then at m=50 and 100, it obtained

comparable results. Eventually, in the third and fourth scenarios, V(λ)=0.5 represents the weak priors that is the closest to strong priors.

Setting 2 V(λ)=0.05 and V(λ)=0.1, as they were the most consistent with the criteria presented. They provide a high coverage in all scenarios, while having the best MSE in the first and the second scenarios.

Setting 3 V(λ)=0.01, as it was the most balanced prior with regard to MSE and coverage rates. In the first two scenarios, they had the same results as V(λ)=0.05 and V(λ)=0.1, while obtaining lower MSE in the third and the fourth scenario.

Setting 4 V(λ)=5 and V(λ)=10, as they had the advantage to provide high coverage in all scenarios, while having the lowest MSE in scenario one. In the three other scenarios, they were either very close or equal to MSE of strong priors. Regarding strong priors, if their results were overall the best in the scenarios two, three and four, they were highly penalizing in the first one in terms of coverage.

In the end, the relevant levels of information to introduce in the settings were always part of the weak priors group. However, in three settings, we noted that it corresponded to the more informative weak priors, i.e. the more restricted in terms of parameter variations.

In the second part of the final objective, we wanted to determine if a strong prior represented a relevant decision when eliciting a prior. We analysed this in terms of risks, i.e. if trialists select a strong prior, what are the consequences on prediction accuracy if it reveals to be off target? Regarding, the first scenarios, strong priors were never relevant as their results were always the worst Moreover, the stronger the prior, the poorer were its MSE and coverage rates. Regarding, the second and third scenarios, depending on the settings, we came to different conclusions explained below:

- In the first setting, with a difference of 6 months between T_{real} and $T_{expected}$, results were mitigated, as we could not favor strong or weak priors. It all summed up to a decision between being more accurate at predicting the real recruitment values, but with less chance to contain those values or the opposite. Nonetheless, except for V(λ)=0.01, the stronger prior, the risks of using more informative priors were not that high, since the

loss in coverage compared to weak priors was small. Therefore, even with a prior 6 months off target, it was relevant to consider their introduction in the model, but not the strongest value, i.e. $V(\lambda)=0.01$.

- In the second setting, when the planned duration error was 3 months, they were bad consequences in using strong priors, as they obtained low coverage. Hence, in those settings, being off target translated into higher risks for strong priors than it was in the first setting.
- In the third setting, when the planned duration error was 3 months, the situation was the same as in the first setting with an error of 6. Indeed, we could neither favor the weak or the strong priors.
- In the fourth setting, as strong priors obtained the best MSE and high coverage with a 6 months planned duration error, they involved even less risks than in the first setting. Hence, it was also relevant to consider them when eliciting the prior.

Therefore, when the prior was entirely off from the target, in all settings the most informative priors represented huge risks. However, when the planned duration error was smaller, the settings provided different results. Regarding the planned duration errors of 6 months, it had a negative impact on the use of strong priors in two settings (the second and the third), whereas for the other two (the first and the fourth), strong priors represented relevant alternatives. Regarding the planned duration errors of 3 months, it had a negative impact on strong priors only in the second setting. Nonetheless, usually the stronger was the prior, the higher were the risks, as they were most impacted by an increase in planned duration error.

4 Discussion

In the previous chapter, we presented the results of several simulation studies. At the end of the section, we gave our main observations on the results in relation with the objectives of the thesis. However, these observations are highly specific to the settings used. The aim of this chapter is, thus, to present the main findings of the simulation studies in a more general perspective, compare them to the scientific literature and discuss the limitations of this work as well as its implications. Eventually, we will provide resulting areas for future research.

The first objective of this study aimed at observing the effects of planned duration errors on strong and weak priors. In every setting, results provided the same conclusion. That (i) the performances of strong and weak priors differ, and (ii) this difference varies depending on the difference between real and expected durations. More informative priors obtain the most accurate predictions when accrual is on target, whereas less informative priors are more effective when accrual is entirely off from the target (12 months). Thus, it was relevant to keep on exploring the introduction of strong priors and determine at which difference of months we would observe this shift in performance between strong and weak priors.

The second objective aimed at understanding when $T_{expected}$ could be considered as close to T_{real}. The first objective gave us some hints on the question, by providing limits on the definition of "closeness". We know the correct value is situated somewhere between 0 and 11 months. To narrow this interval, we used two scenarios where the expected durations were situated at 3 and 6 months from the real one. Unfortunately, we did not find consistent observations across the settings. Hence, we believe that there is no unique definition of the term "close" when considering the proximity between real and expected durations at the beginning of a trial. Moreover, this notion seems to be related to the theoretical constants n and $T_{theoric}$, since the intervals were different given a specific setting. For example, when n=500 and $T_{theoric}$=20, the interval was between 0 and 2, whereas when n=48 and $T_{theoric}$=24, the interval was between 7 and 11.

The last objective of the study was split in two parts (i) define the most relevant level of information to be introduced at the beginning of a clinical trial and (ii) assess the relevance of introducing strong prior information at this stage. For both, we focused on the issues of drug availability in clinical trials, and, thus put more emphasis on the ability of priors to cover the

real recruitment values correctly. Hence, a prior was considered relevant if it had a high coverage compared to others given any difference between real and expected durations, while having or being close to the best MSE.

For the first part, all the settings were leaning towards the use of weaker priors. They were the most consistent with the two criteria presented above. The problem with strong priors was that prediction accuracy, both in MSE and correct coverage, decreased significantly as the prior moved farther away from the target. Concerning weak priors, we noted that, in three settings, the most relevant ones corresponded to the more restricted in terms of parameter variations. Therefore, we would think that when introducing prior information into a model at the start of a trial, being too broad on the parameter space could not be appropriate. If the weakest priors obtained the best results when $T_{expected}$ is at 12 months from T_{real}, they were, however, more penalizing in terms of prediction accuracy in other scenarios. Nonetheless, the second setting showed a contrary position, as it was the only one in which they represented the best alternatives. Overall, we could not confirm whether or not the weakest priors could be introduced at the start of a trial, but we could conclude that, given any types of theoretical rates for a clinical trial, the relevant level of information to introduce seems to be a weak prior.

Eventually, for the second part, we evaluated strong priors in terms of risks, which provided two main findings:

- As for the closeness, the risks seem to be dependent on the settings. Indeed, planned duration errors of 3 and 6 months would not translate in the same risks depending on the values of n and $T_{theoric}$. As an example, there was more risks using strong priors with a planned duration of 6 months in the third setting than in the first one.
- We found that, the stronger the prior, the higher the risk in any given setting. The issue with priors that are over-restrictive on the parameter space, i.e. very little variation around expected value, is that their coverage rapidly decreases in predictions time points and the predictions do not fit the data fast enough as the difference between T_{real} and $T_{expected}$ gets larger. The latter argument is motivated by the relative steadiness of the MSE along m values in the three scenarios.

Those results suggest that situations exist where it is relevant to consider the introduction of strong priors in the model at the beginning of a trial. However, introducing over-restrictive priors still involve great risk to incur. Therefore, in our view, doing so is not to be considered if one wants to limit risk.

Now that we have presented the main findings, we wanted to relate them to the scientific literature on the topic. We identified two articles that used the same model of recruitment to explore the subject of prior choice: Gajewski & al. (2008) and Jiang & al. (2014). The first one was not relevant, as it was mainly presenting the model, but the second one provided a simulation study that could be linked to our results. Given several scenarios, they were trying to identify the best level of information in a prior to predict the real duration of the study (in days). More interestingly, they were specifically looking for priors that recognize early if a trial is off schedule. The only difference lied in their simulation procedure, they kept the same prior mean, that was on target, in each scenario while varying the simulated data. Hence, they considered only the case of one expected rate.

Overall, their conclusions from the simulation study were comparable to ours. They observed that strong informative priors work better when the accrual is on target, whereas weak priors are more efficient when the accrual is entirely off target. Nevertheless, in the end, they could not settle on a best alternative among priors, they were balanced between "*the greater precision provided by the strong priors and the ability of the weak priors to recognize more rapidly problems with slow accrual*" (Jiang & al., 2014, p.626). We had the same kind of trade-off when eliciting the most relevant priors between being more accurate at predicting the real recruitment or having the higher coverage, but we put more emphasis on the second one, which simplified the decision.

Something that Jiang & al. (2014) took in account in their work that we did not consider is that accrual rate are not constant, thus, they change over time in a clinical trial. This represents the first limitation of our study because we assumed that accrual rates were constant over time. However, this constancy is matter of debate in the scientific literature. Indeed, while Zhang & Long (2010) do believe that "*in most real trial situations, the constant accrual rate assumption does not hold*" (p. 650), Gajewski & al. (2012) found evidence of the contrary. The main advantage of a constant rate is the use of a simpler model. The latter offers some benefit compared to more complex models developed in the literature (Gajewski & al., 2012):

- Easier to specify a prior distribution, while complex models involve more parameters, requiring the elicitation of a prior for each one.
- Simple models perform better when the data are limited (small portion of subjects have arrived in the trial), because of the lesser number of parameters to estimate. It is

particularly relevant when the field of study concerns the early part of a trial, where limited data occurs.

- Simple models offer analytical solutions for the posterior distributions. It eases the work of researchers, which do not have to rely on complex and time-consuming algorithms.

Those points correspond to the goal of this thesis. We wanted to design a model where the information in the prior could be easily introduced and clearly understandable. Moreover, we wished to study the early part of a trial, when there are only few data to adjust the prior and make prediction. Thus, we thought that the use of a constant rate was justified, although possibly unrealistic.

A second limitation of this work concerned the independence and identical distribution assumption on waiting times. Zhang & Long (2010) stated that "*the daily accrual is not necessarily i.i.d.*" (p. 650) referring to the approach in Gajewski & al. (2008). To illustrate this, they showed evidence of patterns in accrual for a specific trial. However, it corresponded to a multi-centric trial, involving several medical sites, which is not the case in this work. Gajewski & al. (2012) tested the i.i.d. assumption for waiting times of a specific trial in a unique center, and found no evidence suggesting it would not hold. Thus, we considered that, in this setting, the assumption was reasonable.

The third limitation of this work is the number of centers considered. Real trials generally include several medical centers, which involves that recruitment does not happen in one single place, but rather in several ones situated in different regions, countries and continents. This has many implications for firms. First, it complicates the drug distribution planning, since they need to anticipate future subjects' arrival in more than one place. Secondly, they need to monitor the recruitment on the whole, as well as locally, in each center. Hence, they need quantitative tools that account for the different rates of recruitment at each medical site.

For this thesis, it also has major implications, because it limits its scope. In the case of a multi-centric trial, this study is less relevant to evaluate the performances of the trial on the whole, as most of our assumptions do not hold anymore. Firstly, it does not seem appropriate to consider a single general accrual rate for all subjects' arrivals considering each center recruits at a different rate. We would need to account for this variation of rates in the model, which is obviously not the case. Secondly, as the centers have different accrual rates, the assumptions

that the arrival times between all patients in the trial are i.i.d. does not fit anymore. It would not seem correct to consider that waiting times in a high rate recruiting center are identically distributed compared to a center with a much slower recruitment. Thirdly, as highlighted by Zhang & Long (2010), accrual rates are non-constant in multi-centric trials. Accordingly, while not being relevant for multi-centric trial as a whole, our findings could still be useful when assessing the performances center-by-center. By focusing on one center at a time, the model and assumptions are valid as we are in the same situation as this study. The idea would be to monitor and predict recruitment locally, i.e. in each center, then pool the results to verify that the overall recruitment still goes according to plan. This would account for center-to-center variability in rates and would allow to verify that each center recruits as expected and to determine the quantity of drugs to be delivered on each site. One of the drawbacks would be that the more centers in the trials, the more models it would involve, which would, in return, increase the time needed to obtain the results. It is interesting to note that, this center-to-center evaluation of performances was considered by Jiang & al (2016), whom precisely use this model to monitor that centers were not significantly slow or inefficient in enrollment.

Eventually, the last two limitations of this work concern the way we conducted the simulation studies. Firstly, we only used a limited number of settings, which restrict the implications of our findings. Regarding the relations involving the settings and the introduction of strong priors, we could effectively observe there was some link. Unfortunately, we could not observe whether this was mainly due to the target sample sizes, the theoretical durations chosen or both. Thus, it would need further simulation studies to deepen the understanding of this relation. To achieve this, we provided in **Annexe 1** an R function that performs the same analysis we conducted. It can be easily modified to include different distributions in the prior or the likelihood. Nonetheless, if the small number of settings limited the generalization of the results, we could still observe relevant patterns from our observations.

The second limitation about simulation studies concerns the decision to work with fixed variance values over scenarios. It may have been more appropriate to use fixed coefficients of variation, that provide the advantage of accounting for the uncertainty around the parameter of interest relative to its mean value. Hence, we could have obtained consistency across scenarios and settings, i.e. the same variations around expected values. With fixed variances, it was impossible to associate levels of information from one setting to another. Thus, conclusions on

the best priors were highly specific and could not be compared. Nonetheless, we believe fixed variances were consistent to represent overall weak and strong prior information.

In summary, the limitations presented above restrained the implications of this thesis for real clinical trials. Indeed, most of the findings cannot be linked to the overall monitoring of multicentre trials, as they are not very adapted statistically with inadequate assumptions and modeling. However, considering the case of recruitment in one center, either mono-centric trials or center-by-center in multi-centric trials, we explained above why the findings of this thesis are relevant. Therefore, we can explore their implications and provide some advice to the trial teams.

The main implication concerns the prior choice at the beginning of a study when considering the modeling of an ongoing recruitment. We wanted to determine whether a relevant level of information to introduce into a model in any given clinical trials exists. The elements presented together above provide the conclusion that if trialists want to use prior information, we would recommend them to favor weak priors. Obviously, they would not be as accurate as strong priors when on target. However, given the problem of overestimation in recruitment duration, it would represent a more rational decision, since weak priors involve less risks when off target. Furthermore, we know that drug availability in the medical center represents a crucial determinant of clinical trials. Thus, we would like to favor priors that ensure high coverage of real recruitment data. Weak priors would maybe have less accurate predictions, but would be more robust in planning drug distributions and anticipating shortage.

A second implication concerns the relevance of introducing strong prior information into a model at the beginning of a study and the risks associated with it. Following the observations made above, trialists should know that situations where it seems relevant to introduce strong priors exist. However, two elements are to be considered:

- The risks of using strong priors seem to vary depending on the situations, but it is clear that the stronger is the prior, the higher are the risks.
- Trialists should never be overconfident in their prior by over-restricting the parameter space. As we saw, very strong priors are highly sensitive to planned duration errors, specifically in terms of correct coverage. Thus, they should not be used.

To conclude this discussion, in the lights of all the limitations of this work, we provided areas of improvement to deepen this research. They were split into three different directions: (i) the case of recruitment in one center, (ii) the case of multi-centric trials and (iii) the priors and the way they are constructed. Firstly, concerning the recruitment in one center, we found three main resulting areas of improvement:

- We could extend this study by observing if the conclusions hold for real clinical trials, but this would require the access to real data.
- We could extend this study by considering the same model, but in the case of non-constant accrual, as it was done in Jiang & al. (2014).
- We could extend this study by considering more complex models, involving two or more parameters as the Weibull or Log-normal, which are commonly used by Pharmalex.

Secondly, considering multi-centric trials, we found three main areas of research:

- We could extend this study by analyzing the effects of information levels on several centers in the same trial. Put different priors (strong/weak) on the different centers and evaluate the impacts on the global monitoring of a trial.
- We could extend this study by considering a more complex exponential model accounting for recruitment rates in each medical center. In this case, we think of a frailty model that adjusts for center-to-center variability in recruitment rates, which is also commonly used by Pharmalex.
- We could also consider the case of more complex models that accounts for center-to-center variability in recruitment rates.

Thirdly, considering the priors and the way they are constructed, we found several main areas of research:

- As we highlighted, in the end, the choice on the prior information will depend on the risks one is ready to take if the priors reveal to be off target. Following this idea, we could investigate adaptive priors which in some way would take this risk into account and change the prior given the data observed. We would not be limited to a decision between strong and weak priors and could be more confident in our predictions. This topic was notably considered in Jiang & al. (2014), but their priors did not reach better results compared to more simple weak and strong priors. Hence, it represents a main area of research.

- Following the same idea, we could simply explore and evaluate other methods to build priors in those models like in Gajewski & al (2008) or Jiang & al. (2014).
- We could extend the analysis by considering other measures to assess the relevance of levels of information like in Jiang & al. (2014). They considered the ability of priors to predict accurately the real duration of a study, as well as recognizing early if a trial was off schedule.
- We could conduct a sensitivity analysis to deepen the definition of closeness between expected and real duration given several theoretical settings.
- We could evaluate the inverse effect, when trialists overestimate the study duration.
- We could assess the appropriate number of subjects from which it is no longer relevant to add information in the prior.
- We could further explore the relations between the use of strong priors and the target sample sizes, the theoretical durations chosen or both. For example, by considering a fixed theoretical duration and several n values, we could observe if the performance among levels of information change significantly. The same could be done with fixed n and several $T_{theoric}$.

5 Conclusion

Designing efficient drug distribution planning represents a critical aspect of clinical trials, since inadequate supply, i.e. problems of shortage on medical sites or on the contrary, oversupply, as well as overproduction can be terribly costly for pharmaceutical firms (Fleischhacker & Zao, 2011). Hence, relevant quantitative tools must be developed to optimize the production and the distribution to the different centers. The latter should help companies in providing the most appropriate quantity of drugs to ensure that treatments are always available on medical centers, while minimizing the waste and costs.

In this context, the ability to model and predict patient recruitment accurately is crucial. Accurate prediction methods provide firms a complete information on the plausible arrival dates for subjects in the study. Moreover, it allows them to compute confidence region, which would help them anticipate any issue concerning enrollment and prepare robust drug distribution planning. If several methods have been developed to achieve this, either before or during the recruitment (Gkioni & al., 2019), some issues still arise. One of them concerns the predictions of accrual at the beginning of the recruitment, when few data are available, and represented the topic of research of this thesis.

More specifically, we focused this work on a specific prediction method involving a Bayesian approach. The objective was to assess the relevance of introducing strong prior information when monitoring the start of a trial and, more generally, determining what was the most relevant level of information to be introduced at this stage. The first chapter of this thesis aimed at introducing the underlying theory that supported the empirical study, by reviewing the main characteristics of the Bayesian statistics as well as the Markov Chain Monte Carlo approach. In the second chapter, we conducted simulation studies to explore the objectives. Then, as a third chapter, we discussed the findings, compared them to scientific literature, presented their limitations, derived their main implications in real clinical trials and, eventually, provided some areas of improvement.

We observed two main findings in this thesis. Firstly, weak priors represent overall the most relevant level of information to be introduced into a prior. Indeed, to avoid drug shortage in the medical center, they represent the best alternative as they ensure a high coverage given any planned duration error considered. Regarding predictions accuracy, if they are less efficient

than strong priors when on target, they represent a more adaptive solution that can quickly accommodate with an error in planned duration. This make even more sense when we know that trialists tend to underestimate their recruitment duration. Thus, they represent the most relevant choice among priors, in order to build robust drug distribution planning. Secondly, the use of strong priors is situation-specific, and, that they overall do not correspond to the most relevant level of information to be introduced into a prior. If they provide the best predictions accuracy and coverage when on target, they generally lose this advantage as the planned duration error increases.

Regarding the implications in the literature, we identified one work that explored influence of the prior choice on the same type of model. Their objectives were not exactly the same as they focused on the ability of those priors to predict accurately trials duration and be able to recognize early the accrual rates are off target. Moreover, their simulation procedure was slightly different as involving non-constant accrual rates. In the end, their observations from the results did correspond to ours, the only difference being that they could not make an obvious choice among priors.

Regarding the implications of this work in clinical trials, we faced two situations. (i) Concerning multi-centric trials, the assumptions and the research were not adapted to fit with the global recruitment monitoring and, thus, the findings could not be related to them. The main reason is simply that we assumed a unique accrual rate, which was not relevant, as multi-centric trials involve several medical centers with different accrual rates. We were not able to account for this variability in accrual rates with the model used. (ii) Regarding mono-centric trials or center-to-center evaluation in multi-centric trials, we found relevant implications for the modeling of subjects' recruitment. The main one concerns the inclusion of strong priors to improve predictions at the start of a trial. As presented above, we found evidence that situations exist where they would be relevant. However, they still represent a risk as their coverage of real recruitment values decrease with an increase of the planned duration error. Therefore, we believe weak priors represent the safest decision when introducing information into a model, providing the best compromise in terms of errors in prediction and correct coverage.

Many scientific research focused on developing relevant methods for predicting accrual at the design stage of trials, for both mono and multi-centric, but hardly assessed their relevance at their beginning, when the information provided by the data is uncertain. Yet, it corresponds to

the stage where the methods need to be the most robust and efficient as the impact of the predictions are highly significant on the remaining of the trial. Indeed, trialists need to verify if their design stage planning was accurate, while providing the necessary change to optimize the drugs production and distribution planning. However, as highlighted by Zhang and Long (2010, p. 657): "*an early projection is more likely to have a wide CI, especially when there does not exist good prior information. Projection with wide CIs may provide limited guidance in practice*". This stresses the development of methods to improve this guidance. This study made a first step towards this direction. Therefore, we would encourage people to keep working on this topic, for example, by exploring the issues of multi-centric trials, non-constant accrual rates, more complex models or the construction of robust priors that account for risks associated with a level of information.

6 Bibliography

- Barnard, D. K., Dent, L., & Cook, A. (2010). A systematic review of models to predict recruitment to multicentre clinical trials. *BMC Medical Research Methodology*, *10*(63).

- Boldstad, W.M. (2007) *Introduction to Bayesian Statistics* (2nd Ed.). Wiley.

- Chen, Y., Mockus, L., Orcun, S., & Reklaitis, G. V. (2010). Simulation-based optimization approach to clinicaltrial supply chain management. In S. Pierucci, & G. Buzzi Ferraris (Eds.). *Computer Aided Chemical Engineering*, *28*, 145–150.

- Deng, Y., Zhang, X. and Long, Q. (2014). Bayesian modeling and prediction of accrual in multi-regional clinical trials. *Statistical Methods in Medical Research*, *26*(2), pp.752-765.

- Enders, C. (2010). *Applied Missing Data Analysis*. New York: Guilford Publications.

- Fleischhacker, A. and Zhao, Y. (2011). Planning for demand failure: A dynamic lot size model for clinical trial supply chains. *European Journal of Operational Research*, *211*(3), 496-506.

- Gajewski, B., Simon, S., Carlson, S. (2008). Predicting accrual in clinical trials with Bayesian posterior predictive distributions, *Statistics in Medicine*, *27*, 2328–2340

- Gajewski, B., Simon, S., & Carlson, S. (2012). On the Existence of Constant Accrual Rates in Clinical Trials and Direction for Future Research. *International Journal Of Statistics And Probability*, *1*(2). doi: 10.5539/ijsp.v1n2p43

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd Ed.). Boca Raton, FL: CRC Press.

- Gkioni, E., Rius, R., Dodd, S., & Gamble, C. (2019). A systematic review describes models for recruitment prediction at the design stage of a clinical trial. *Journal Of Clinical Epidemiology*, *115*, 141-149. doi: 10.1016/j.jclinepi.2019.07.002

- Healy, P., Galvin, S., Williamson, P., Treweek, S., Whiting, C., & Maeso, B. et al. (2018). Identifying trial recruitment uncertainties using a James Lind Alliance Priority Setting Partnership – the PRioRiTy (Prioritising Recruitment in Randomised Trials) study. *Trials*, *19*(1). doi: 10.1186/s13063-018-2544-4

- Jackman, S. (2009) *Bayesian Analysis for the Social Sciences*. Wiley

- Jiang, Y., Simon, S., Mayo, M., & Gajewski, B. (2014). Modeling and validating Bayesian accrual models on clinical data and simulations using adaptive priors. *Statistics In Medicine*, *34*(4), 613-629. doi: 10.1002/sim.6359

- Jiang, Y., Guarino, P., Ma, S., Simon, S., Mayo, M., Raghavan, R., & Gajewski, B. (2016). Bayesian accrual prediction for interim review of clinical studies: open source R package and smartphone application. *Trials*, *17*(1). doi: 10.1186/s13063-016-1457-3

- King, R. (2010). *Bayesian analysis for population ecology*. Boca Raton: CRC Press.

- Kruschke, J. (2015). *Doing Bayesian Data Analysis* (2nd Ed.). Boston: Academic Press.

- Lan, Y., Tang, G., & Heitjan, D. (2018). Statistical modeling and prediction of clinical trial recruitment. *Statistics In Medicine*, *38*(6), 945-955. doi: 10.1002/sim.8036

- Mijoule, G., Savy, S., & Savy, N. (2012). Models for patients' recruitment in clinical trials and sensitivity analysis. *Statistics In Medicine*, *31*(16), 1655-1674. doi: 10.1002/sim.4495

- Pemberton, V., Evans, F., Gulin, J., Rosenberg, E., Addou, E., & Burns, K. et al. (2018). Performance and predictors of recruitment success in National Heart, Lung, and Blood

Institute's cardiovascular clinical trials. *Clinical Trials*, *15*(5), 444-451. doi: 10.1177/1740774518792271

- Peterson, M., Byrom, B., Dowlman, N. and McEntegart, D. (2004). Optimizing clinical trial supply requirements: simulation of computer-controlled supply chain management. *Clinical Trials: Journal of the Society for Clinical Trials*, *1*(4), pp.399-412.

- Robert, C. (2001). *The Bayesian Choice* (2nd Edition). Springer.

- Sully, B., Julious, S., & Nicholl, J. (2013). A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. *Trials*, *14*(1), 166. doi: 10.1186/1745-6215-14-166

- Zhang, X., & Long, Q. (2010). Stochastic modeling and prediction for accrual in clinical trials. *Statistics in Medicine*, *29*, 649-658

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**