Novel Bayesian approaches for the planning and monitoring phases of clinical studies

Armando Turchetta

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec
August 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

I dedicate this thesis to my Mom and Dad.

Acknowledgements

This thesis represents the conclusion of a five-year journey which I could not have taken without the help, support, and guidance from many people I am thankful to.

First and foremost, I want to thank my co-supervisors, Dr. Erica E.M. Moodie and Dr. David A. Stephens, for everything they have done for me during the entirety of my PhD. I am deeply grateful to Erica for being an outstanding and enthusiastic mentor from before she even became my supervisor, and for her relentless encouragement and support, especially during the first year of the pandemic and other periods when I did not feel good about the state of my research. I am greatly indebted to David for his guidance, invaluable inputs, and for sharing his knowledge in our many, many meetings.

I want to thank my committee member, Dr. Nicolas Savy, for allowing me to travel to his department at the Toulouse Mathematics Institute for a research visit and for being incredibly helpful with his guidance and kindness during my stay in France and beyond.

I would also like to thank the many people who helped me acquire the data used in this thesis, including Dr. Sylvie D. Lambert, Manon de Raad, Dr. Marina B. Klein, Dr. Michael Kramer, and the HIV Vaccine Trials Network team.

I wish to acknowledge the financial support I received throughout my PhD from my supervisors, the Department of Epidemiology, Biostatistics, and Occupational Health, and the Fonds de recherche du Québec – Nature et technologies.

There have been several professors in the Department of Epidemiology, Biostatistics, and Occupational Health who have been an important source of influence for me. I would like to particularly thank Dr. James A. Hanley, Dr. Andrea Benedetti, Dr. Shirin Golchi, and Dr. Alexandra M. Schmidt. I am also grateful to the department's administrative staff, who have always been very helpful throughout my studies at McGill, from the acceptance letter to the thesis defense, especially Katherine Hayden, André Yves Gagnon, and Dolores Coleto.

I am grateful to all my Montreal friends for their support and for making my life so much more enjoyable and fun during these five challenging years. In particular, I would like to thank Giulio, Daniel, Vanessa, Victoire, Larry, and Charlie. I wish to thank my girlfriend, Stephanie, for her love and patience throughout the last year of my PhD, especially during the writing of this thesis.

There have been several people who have had a crucial role in helping me get to McGill in the first place. I am indebted to my master's thesis advisors, Dr. Francesca Dominci and Dr. Luca Tardella, who encouraged me to pursue a doctoral degree and supported my application. I would like to thank my friend Sara, who was very supportive of my decision to start a PhD and helped with the applications. I also wish to thank my Aunt Giovanna for instilling in me the idea of an education in statistics when I was in middle school.

Lastly, an enormous thank you goes to my family: my many sisters – Laura, Beatrice, and Valentina – and especially my parents, Luigi and Giuliana, for... Well, just about everything, from emphasizing the value of education since I was a child to being the all-around best parents anyone could have ever asked for.

Preface

This manuscript-based PhD thesis contains six chapters: an introduction, a literature review, three stand-alone original manuscripts which correspond to Chapters 3-5, and a conclusion which includes a comprehensive discussion of the findings and implications of the thesis.

The introduction, literature review, and conclusion contained in Chapters 1, 2, and 6 were written by Armando Turchetta (AT) and revised by Erica E.M. Moodie (EEMM) and David A. Stephens (DAS), who also provided guidance for their conception.

The methodological work in the first manuscript presented in Chapter 3 was conceptualized in discussions between AT, EEMM, and DAS. The methodological derivations, simulation study, and data analysis were performed by AT under the supervision and guidance of EEMM and DAS. The data for the case study analyzed in the manuscript were provided by Sylvie D. Lambert (SDL). The manuscript was written by AT and revised by EEMM, DAS, and SDL.

The methodological work developed in the second manuscript (Chapter 4) was originally conceptualized through discussions between AT, EEMM, and DAS, and further elaborated with the help of Nicolas Savy (NS). The methodological derivations, simulation study, and data analysis were carried out by AT with guidance from NS, EEMM, and DAS. The data analyzed in the case study presented in the manuscript were provided by Marina B. Klein (MBK). The manuscript was written by AT and revised by NS, EEMM, DAS, and MBK.

The ideas for the third manuscript (Chapter 5) were conceptualized by AT, EEMM, and DAS. AT performed the analysis and developed the R package under the supervision and guidance of EEMM and DAS. The data analyzed in the main manuscript were provided by the HIV Vaccine Trials Network (HVTN). The case study illustrated in the supplementary material of this manuscript (Appendix C) is based on data provided by Michael Kramer. The manuscript was written by AT and revised by EEMM, DAS, and NS. The teams at

HVTN and Janssen Pharmaceuticals (which included Zoe Moodie, Janine van Duijn, and Wouter Willems) provided comments and suggestions. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases, the National Institutes of Health, or by Janssen Pharmaceuticals.

Abstract

Randomized clinical trials are of paramount importance in medical research and are particularly valuable in understanding causal relationships, as confounding is removed through the randomization process. With new designs being developed and the use of multicenter studies increasingly common, trials are growing in complexity, and their associated cost has been increasing year on year for the past two decades. Unfortunately, many trials fail, perhaps through poor planning (underestimating sample size needs) or the inability to meet accrual targets. This thesis considers aspects of these challenges in planning and monitoring, developing new Bayesian approaches to sample size calculations for a multi-stage randomization design and accrual monitoring for multicenter studies. A novel trial design that has earned the spotlight in the growing field of precision medicine is the sequential multiple assignment randomized trial (SMART). Within this design, patients are randomized at two or more key treatment stages accounting for a small set of characteristics or responses to previous interventions. This structure allows for the development and comparison of adaptive treatment strategies. Most of the primary analyses performed on SMARTs are based on the comparison of two means or strategies, and while the frequentist sample size formulae are similar to traditional randomized controlled trials (RCTs), their estimation relies on additional assumptions. In the first manuscript, I developed a more robust sample size methodology in the Bayesian framework by adapting the 'two priors' approach to the SMART design while incorporating estimates of the variance components and their uncertainty from pilot studies, resulting in a methodology that relies on fewer assumptions, is more robust to model misspecification, and allows for the incorporation of pre-trial knowledge. The performance of this approach is compared to the frequentist formulae in a simulation study. Its properties are further displayed in a case study where I used data from a SMART pilot to estimate the sample size of its full-scale version. In the second manuscript, I turn my attention from the planning to the monitoring phase, developing a novel approach to forecasting enrollments in multicenter studies applicable to both cohort and trial designs. The forecasting of recruitments is a topic of rapidly increasing interest, however, most models used in practice are either deterministic or rely on often unrealistic assumptions, such as the constant recruitment intensity over time. The most popular methodology that belongs to the second group is the Poisson-Gamma (PG) model. I extended this methodology by allowing the enrollment rates to vary with time after the opening of recruitment centers up to a stabilization point. I illustrate the accuracy of this methodology compared to the standard PG model in a simulation study and by forecasting the enrollments in the Canadian Co-infection Cohort study. The third manuscript aims to further validate the proposed recruitment model on data from randomized trials and offer a practical guide for its use. One of the main hurdles to the adoption of statistical models to forecast enrollments in practice lies in the difficulty of their implementation. In this manuscript, I outline how to implement the time-dependent PG model to predict the recruitment process via the newly developed `tPG` R package. The model is further validated on the recruitment data from two HIV trials.

Abrégé

Les essais cliniques randomisés sont d'une importance capitale dans la recherche médicale et sont notamment utiles à la compréhension des relations causales, car la randomisation élimine les facteurs de confusion. Avec le développement de nouvelles planifications et l'utilisation plus courante d'études multicentriques, les essais cliniques deviennent de plus en plus complexes et leur coût augmente tous les ans, depuis une vingtaine d'années. Cependant, de nombreux essais cliniques échouent, peut-être en raison d'une mauvaise planification (sous-estimation de la taille d'échantillon nécessaire) ou de l'incapacité à atteindre leurs objectifs de recrutement. Cette thèse examine certains aspects de ces défis de planification et de suivi, en développant de nouvelles approches bayésiennes au calcul de taille d'échantillon pour un plan à randomisation en plusieurs étapes et pour le suivi du recrutement dans le cadre d'études multicentriques. L'essai randomisé séquentiel à assignations multiples (« SMART ») est un nouveau type d'essai clinique qui s'est imposé dans le domaine en plein essor qu'est la médecine de précision. Dans ce cas, les patients sont randomisés à deux ou plusieurs étapes clés du traitement, selon un petit ensemble de caractéristiques ou de réponses à des interventions passées. Cette structure permet le développement et la comparaison de stratégies de traitement adaptatives. La plupart des analyses primaires effectuées sur les SMART sont basées sur la comparaison de deux moyennes ou deux stratégies; et bien que les formules fréquentistes de tailles d'échantillon soient similaires aux habituels essais contrôlés randomisés (ECR), leur estimation nécessite des hypothèses supplémentaires. Dans le premier manuscrit, je développe une méthodologie bayésienne de calcul de tailles d'échantillon plus robuste en adaptant l'approche des 'deux distributions a priori' au plan SMART. L'inclusion de composantes de la variance, ainsi que leur incertitude, estimées à partir d'études pilotes, résulte en une méthodologie basée sur moins d'hypothèses, qui est plus robuste à une mauvaise spécification du modèle, et qui permet d'incorporer des connaissances préalables. L'efficacité de cette approche est comparée aux formules fréquentistes dans une étude par simulations. Ses propriétés sont aussi démontrées dans une étude de cas où j'utilise les données d'un projet pilote SMART pour estimer la taille d'échantillon de sa version à grande échelle. Dans le second manuscrit, mon intérêt passe de la phase de planification à celle de suivi. Je développe une nouvelle approche de prévision des recrutements aux études multicentriques, applicable aux études de cohorte et aux essais cliniques. La prévision des recrutements est un sujet en croissance, mais la plupart des modèles utilisés dans la pratique sont déterministes ou reposent sur des hypothèses souvent irréalistes (par ex : une intensité de recrutement constante au fil du temps). Dans ce dernier cas, la méthodologie la plus commune est le modèle Poisson-Gamma (PG). Je développe cette méthode en permettant aux taux de recrutement d'évoluer dans le temps après l'ouverture des centres de recrutement, et jusqu'à un point de stabilisation. Je montre la précision de cette approche par rapport au modèle PG standard dans une étude par simulations, ainsi qu'en prévoyant le recrutement à l'étude Canadian Co-infection Cohort. Le troisième manuscrit vise à valider davantage le modèle de recrutement proposé sur des données provenant d'essais randomisés et à offrir un guide pratique pour son utilisation. En pratique, l'un des principaux obstacles à l'adoption de modèles statistiques pour la prévision des recrutements réside dans la difficulté de leur mise en œuvre. Dans ce manuscrit, j'explique comment utiliser le modèle PG temporel pour prédire le processus de recrutement via la nouvelle librairie R `tPG`. Le modèle est en outre validé sur les données de recrutement de deux essais cliniques sur le VIH.

Table of contents

List of Tables

List of Figures

Abbreviations

ADHD Attention-deficit/hyperactivity disorder

AIDS Acquired immunodeficiency syndrome

ATS Adaptive treatment strategy

BCAWS Biased coin adaptive within-subject

BIC Bayesian Information Criterion

BOBYQA Bound Optimization BY Quadratic Approximation

CCC Canadian Co-infection Cohort

CDF Cumulative distribution function

COVID-19 Coronavirus disease 2019

CR Coverage rate

CrI Credible interval

DASS Depression Anxiety and Stress Scale

DTR Dynamic treatment regime

HAART Highly active antiretroviral therapy

HCV Hepatitis C virus

HIV Human immunodeficiency virus

HVTN HIV Vaccine Trials Network

MCB Multiple comparisons with the best

MDD Minimal detectable difference

MLE Maximum likelihood estimation

MM Marginal mean

NB Negative binomial

NHPP Nonhomogenous Poisson Process

NIAID National Institute of Allergy and Infectious Disease

NIX Normal-inverse-chi-squared

PDF Probability density function

PG Poisson-Gamma

PROBIT Promotion of Breastfeeding Intervention Trial

RCT Randomized controlled trial

SD Standard deviation

SE Standard error

SMART Sequential multiple assignment randomized trial

SMD Standardized mean difference

tPG Time-dependent Poisson-Gamma

UK United Kingdom

WHO World Health Organization

Chapter 1

Introduction

Randomized controlled trials (RCTs) are widely considered the gold standard for the evaluation and comparison of the efficacy of treatments [Heindel et al., 2022]. The randomization process and prospective data collection remove most sources of bias that affect observational studies. The random allocation of treatments in a well-designed RCT eliminates confounding bias by balancing the measured and unmeasured risk factors across treatment groups, allowing researchers to draw causal conclusions from the outcomes of the study [Hariton and Locascio, 2018]. RCTs can, however, be highly time-consuming and expensive. With new complex designs being developed and the increasing propensity to conduct trials in multiple centers across various countries, the cost of conducting a study from protocol approval to trial report has been steadily growing over the last two decades [Gresham et al., 2020]. Nonetheless, many clinical trials fail to achieve their primary goal of finding a statistically significant clinically meaningful difference between randomization arms. In part, this is a natural consequence of the clinical equipoise that is necessary for the ethical conduct of clinical trials [Freedman, 1987]. However, a significant number of trials are discontinued because of poor recruitment, or they fail to show a significant result because of an inadequate sample size based on flawed assumptions. These are intertwined, long-standing issues in clinical trial conduct that pose ethical concerns regarding the depletion of research resources and patient involvement, as the study participants generally expect that the trial will lead to a societal benefit via the information gained from the study [Fogel, 2018]. This thesis deals with the development of novel statistical tools to aid the planning and monitoring phases of clinical studies within the Bayesian framework, introducing approaches for the sample size determination of an emerging multi-stage trial design and the forecasting of recruitments to monitor accrual in multicenter studies.

One of the most promising trial designs that has emerged in the precision medicine field is the sequential multiple assignment randomized trial (SMART) design [Murphy, 2005b]. Within the precision medicine clinical model, decisions regarding the treatment of individual subjects are tailored to their characteristics, shifting the focus from the traditional 'one size fits all' approach to a more personalized model [Wallace and Moodie, 2014]. The SMART design is based on multiple stages, each representing a clinical decision point: at each stage, the study participants are randomized to the available treatments according to a small set of characteristics, including their response to previous interventions. This adaptive structure, combined with the sequential randomization, allows the trialist to develop and compare treatment strategies, properly known as adaptive treatment strategies (ATSs) or dynamic treatment regimes (DTRs) [Lavori et al., 2000, Murphy et al., 2001]. An adaptive treatment strategy is a sequence of decision rules that inform the clinician regarding the assignment of interventions for the patient concerned, mirroring what is observed in clinical practice. In fact, clinicians often need to choose the next intervention to be assigned to a patient based on their conditions and response to previous treatments, generating treatment paths that are not pre-determined. The SMART design allows formal comparisons of these covariate-adapted strategies in an experimental setting. While the number of SMARTs has been growing in recent years, some operating characteristics are not yet well understood. For instance, the sample size determination for the comparison of two strategies embedded in a SMART has been mainly analyzed in the frequentist setting, often overlooking the plausibility and uncertainty around the additional assumptions and key parameters that need to be posited with respect to a traditional RCT, potentially leading to underpowered studies. In the first manuscript, I extend the frequentist sample size methodology developed in Oetting et al. [2011] for continuous outcomes to the Bayesian framework via the 'two priors' approach [Wang et al., 2002, Sahu and Smith, 2006]. With respect to the frequentist calculations, the resulting methodology relies on fewer assumptions, accounts for uncertainty around the design parameters, and allows for the incorporation of pre-trial knowledge. The approach is validated in a thorough simulation study and is implemented to size a full-scale SMART, namely the Internet-Based Adaptive Stress Management SMART, using data from its completed pilot study.

In the second and third manuscripts, I focus on the monitoring phase of multicenter clinical studies. Forecasting recruitments is a topic with significant practical and financial implications. However, notwithstanding a large body of literature stressing the inadequacy of deterministic models based on the study investigators' expectations to predict enrollments, they are still widely used in practice [Anisimov, 2008, Gkioni et al., 2020]. This approach ignores relevant sources of variability and underestimates the required time to achieve the targeted sample size, as investigators are known to assume overly optimistic recruitment rates. Several approaches to forecast recruitments in multicenter clinical trials have been introduced throughout the years. Notably, Anisimov and Fedorov [2007a, 2007b] proposed the doubly-stochastic Poisson-Gamma (PG) recruitment model. According to this method, participants are enrolled by the various recruitment centers as a Poisson process whose rate parameter originates from a common Gamma distribution. One relevant limitation of this approach is the underlying assumption that recruitment rates remain constant over time, which is often not met in real clinical studies. Additionally, while this model is one of the most popular in the field, its implementation in practice by investigators and statisticians, along with other methodologies, remains scarce [Gkioni et al., 2020].

In the second manuscript, I developed a time-dependent extension of the Poisson-Gamma model which allows the recruitment rates to vary with time following their opening until a plateau or stabilization point. The non-constant section of the recruitment phase is modeled via B-splines to ensure flexibility and capture a wide range of recruitment progressions. The model is validated and compared to the standard Poisson-Gamma model in a simulation study for different shapes of the recruitment curve, including the setting where the constant-rate assumption of the standard PG model is met. The accuracy of the proposed recruitment model is validated in a case study by forecasting the enrollments in the Canadian Co-infection Cohort study, an ongoing prospective observational study that encompasses 19 recruitment sites across Canada [Klein et al., 2010].

In the third manuscript, I further validated the time-dependent PG model on the recruitment data from two HIV trials conducted in multiple Sub-Saharan countries to illustrate the applicability of this approach in randomized controlled trials. Additionally, the manuscript includes a step-by-step tutorial on the newly developed `tPG` R package that implements the functions necessary to estimate, predict, and visualize the recruitment process via the time-dependent PG model, hoping to facilitate the practical implementation of this methodology.

This thesis is manuscript-based and is structured as follows: in Chapter 2, I review the literature of the methodological work that is at the basis of the manuscripts. Chapters 3, 4, and 5 include the three stand-alone manuscripts that represent the core of this thesis, each preceded by a preamble that details the original contributions of each chapter. Chapter 3 was published in *Biometrics*. Chapter 4 was recently published in *Statistics in Medicine*. Chapter 5 will be submitted to a statistical journal soon after the submission of this thesis. Finally, Chapter 6 offers a comprehensive summary of the work developed in this thesis, as well as a discussion of the limitations and possible avenues for future work.

Chapter 2

Literature review

This chapter examines the literature that serves as the foundation of the three manuscripts presented in this thesis, and is divided into four main sections. Section 2.1 offers an introduction to adaptive treatment strategies and the SMART design. In Section 2.2, we discuss the challenges in sizing a SMART study, and review the methodologies that have been proposed for estimating the sample size and their limitations. In Section 2.3, we give a brief overview of the sample size determination from a Bayesian perspective with a particular focus on the 'two priors approach'. Finally, in Section 2.4, we review the existing methods developed to forecast recruitments in clinical trials and their limitations. Please note that in order to use the same notation as that used in the manuscripts, there are slight inconsistencies between the notation used in Sections 2.1-2.3 and Section 2.4. Where these occur, they are explicitly noted to avoid confusion.

2.1 Adaptive treatment strategies and SMARTs

While the majority of randomized controlled trials (RCTs) seek to compare individual treatments or fixed sequences of interventions, in practice, clinicians often need to make decisions on treatment assignments based on the patients' evolving conditions and their history. This is particularly relevant in the treatment of conditions that require a series of interventions, such as chronic diseases, where the clinician has to adapt their approach depending on the patient's time-varying covariates, including their response to previous treatments. This approach generates adaptive sequences of treatments or strategies which, in contrast to most trial designs, are not pre-determined at study start. In observational studies, the estimation of the optimal treatment strategy received some initial attention in Lavori et al. [1994] and Robins [1997], and then its theory was further explored in following publications [Lavori et al., 2000, Murphy et al., 2001, Murphy, 2003, Robins, 2004]. Formally, an adaptive treatment strategy (ATS), also known as dynamic treatment regime (DTR), is a sequence of decision rules regarding the treatment allocation over time. At each decision point, the decision rule takes the subject's time-varying covariates (including previous treatment allocations) as inputs and outputs the next treatment to be assigned. The estimation of the optimal adaptive treatment strategy is a topic that has been extensively studied in observational studies. Some of the methodologies that have been developed for this task include G-estimation [Robins, 2004], Q-learning [Murphy, 2005a, Moodie et al., 2012, Chakraborty and Murphy, 2014], dynamic weighted ordinary least squares [Wallace and Moodie, 2015], and Bayesian machine learning [Murray et al., 2018]. However, the scope of this thesis lies in the experimental context.

Lavori and Dawson [1998] highlighted the need for a statistical framework in the development and comparison of treatment strategies in an experimental setting and reviewed contemporary study designs deployed for these tasks. The authors later introduced a class of randomized designs called 'biased coin adaptive within-subject' (BCAWS) designs [Lavori and Dawson, 2000]. Within this design class, the patient's current condition and response to previous interventions influence the randomization probability to future treatments, so that the treatment patterns are closer to those observed in observational data than in designs that entail a randomization to fixed sequences of treatments. Building on this work, the sequential multiple assignment randomized trial (SMART) design was developed [Murphy, 2005b], and it is currently considered the gold standard trial design for the comparison of ATSs. SMARTs are multi-stage trials where each stage represents a clinical decision point: at each stage, patients are randomized accounting for a small set of characteristics which typically includes their response to previous interventions. Because of their flexibility, various SMARTs can be defined varying the number of stages and treatments available at each stage [Lei et al., 2012]. Figure 2.1 displays two common two-stage SMART designs. Patients are randomized to two first stage interventions and their status is assessed. In the design on the left, subjects who respond to the first treatment continue with the same intervention and non-responders are re-randomized to two second stage treatments, whereas in the design on the right, both responders and non-responders are re-randomized. Note that the second stage treatments for non-responders to different first stage interventions do not need to be different. For example, in the design on the left panel of Figure 2.1, $\{c, d\}$ can, but need not, equal $\{e, f\}$. The same is true for responders to different first stage treatments in the design on the right panel: $\{c, d\}$ may or may not equal $\{g, h\}$.

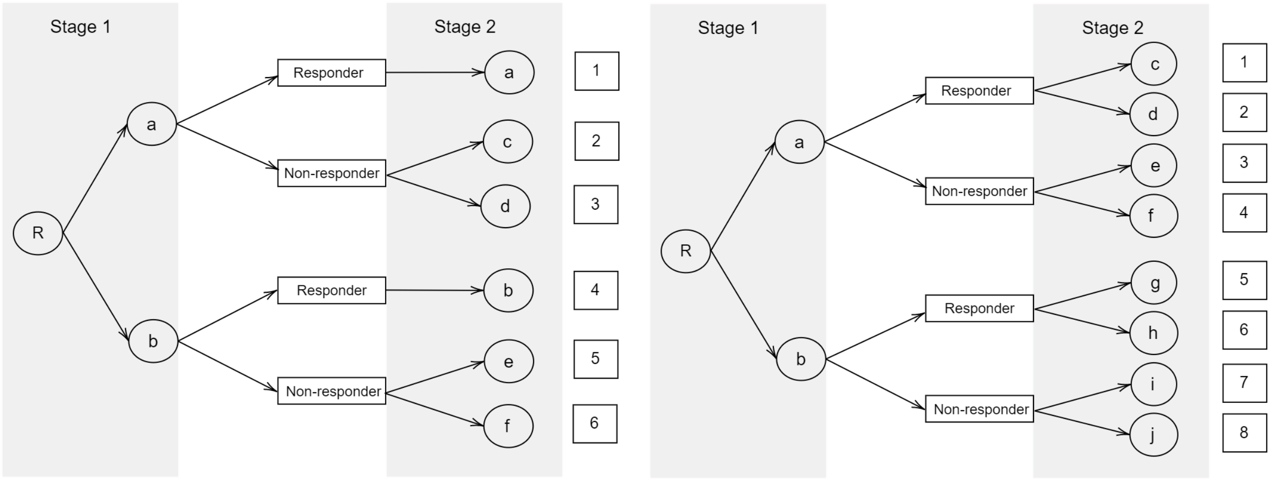

Figure 2.1: Common two-stage SMART designs. The design on the left employs the 'play the winner' approach, whereas the design on the right entails a second randomization for responders to the first stage intervention as well.

Even in the more parsimonious design on the left of Figure 2.1, the SMART design embeds six sequences of treatments (numbered from 1 to 6) and four adaptive treatment strategies of the form 'assign treatment a and, if the subject responds, continue with a, otherwise switch to treatment c'. If responders are re-randomized as well (design on the right), the number of embedded treatment sequences and ATSs each increases to eight, and if a third stage is added, the complexity of the trial increases exponentially. While this highlights the added complexity of a SMART design compared to a standard RCT, the sequential randomization feature of SMARTs and its adaptive structure make them more efficient and effective than other experimental designs that might appear operationally similar, such as multi-arm RCTs where subjects are randomized to fixed sequences of treatments, crossover trials, or factorial trials with a delayed second randomization [Wallace et al., 2016, Heindel et al., 2022]. In fact, while these designs entail the assignment of multiple treatments per patient, not all of them allow for the detection of interactions between treatments (e.g., because of the presence of a wash-out period), and none of them includes an adaptive component. An alternative option consists in stringing together a series of standard RCTs. For example, once an RCT is concluded, additional trials might be conducted for non-responders. Not only is this course of action inefficient, but it might also lead to missing relevant synergistic or delayed effects in a sequence of treatments and to deleterious cohort effects, as subjects who remain in a standard RCT may differ from participants in a SMART study due to, for example, the additional alternatives given to non-responders in a SMART [Murphy et al., 2007]. However, an interesting new design that seeks to incorporate subjects from previous RCTs into a SMART study has been introduced in Liu et al. [2017]. For a more in-depth discussion on the advantages and differences of SMARTs compared to other experimental designs, see Murphy et al. [2007], Lei et al. [2012], Wallace et al. [2016], and Heindel et al. [2022].

Throughout this thesis, we will mainly focus on the design on the left depicted in Figure 2.1, as the two-stage design with the 'play the winner' approach where responders to the first stage interventions continue with the same treatment is the most common and of greater interest for researchers, especially in the field of mental health research [Oetting et al., 2011], where ATSs are commonly sought. Bigirumurame et al. [2022] conducted a systematic review of full-scale SMARTs to evaluate the quality of reporting. Twelve studies were included (along with protocols and methodological papers that did not consider real trials). Of these 12 trials,

- eight studies followed a standard SMART design; among these
  - five followed the same design as that depicted in Figure 2.1 (left);
  - two had a slightly more complex design but still implemented the 'play the winner' strategy;
  - one study did not make clear how the randomization of the second stage treatments depended on the intermediate outcome;
- four studies were multi-stage trials that mostly preceded the popularization of SMARTs; of these, two followed the 'play the winner' strategy.

Similar results were found in the systematic review by Lorenzoni et al. [2023] of SMARTs conducted in oncology. The article included 19 multi-stage trials for which reports of trial's results or protocols were published. While some trials preceded the introduction of SMARTs or were not defined as such in the study design, 17 trials entailed a second randomization only for one subgroup of participants based on the outcomes of the first stage intervention. The 'play the winner' terminology might be misleading in this case, as some trials re-randomized only the patients who showed progress. All trials had two stages.

SMARTs have been implemented in a wide range of fields. Because of their adaptive nature, they are particularly useful in the management of chronic conditions [Chakraborty and Moodie, 2013]. They have been deployed for the management of schizophrenia [Stroup et al., 2003], autism [Kasari et al., 2014], ADHD [Pelham Jr et al., 2016], alcohol dependence [Nahum-Shani et al., 2017], and depression [McCusker et al., 2021], among others. SMARTs have also been implemented in oncology [Kidwell, 2014, Lorenzoni et al., 2023], especially in the treatment of prostate cancer [Wang et al., 2012] or to improve symptom management [Auyeung et al., 2009, Kelleher et al., 2017, Sikorskii et al., 2017]. Additional examples are substance abuse [Murphy et al., 2007], weight loss [Almirall et al., 2014], tobacco cessation [Fu et al., 2017], cannabis use disorder [Stanger et al., 2020], and infectious diseases [Wang and Chakraborty, 2023]. However, although the number of SMARTs is increasing, it remains a novel trial design that, from a statistical perspective, entails unique challenges that are still under investigation. Shortreed et al. [2014] identified five specific challenges in applying imputation methods to SMARTs. These are mainly due to the fact that the timing and number of randomizations are heterogeneous across patients and can depend on their evolving conditions. The authors proposed an ad-hoc time-ordered nested conditional imputation model to tackle these difficulties, but they also highlighted that further work is needed to assess and eventually alleviate some working assumptions. He et al. [2022] pointed out how the missclassification of responders to the first stage intervention leads to inappropriate treatment assignments and can have an impact on the power of the study. Other areas being investigated include adaptive randomization [Cheung et al., 2015, Morciano and Moerbeek, 2021, Wang et al., 2022] and interim monitoring [Manschot et al., 2022, Wu et al., 2023]. In the next section, we discuss some complexities that emerge in the sample size calculations for SMARTs and the existing methodologies.

2.2 Sample size estimation for SMARTs

Since the SMART design embeds multiple treatment sequences and strategies, the sample size determination depends on the definition of the primary research question. There are multiple research questions that can be addressed in a SMART, either as the primary or secondary analysis. The most common are

1. Which is the best initial treatment?

2. Which is the best secondary treatment among non-responders to the first intervention?
3. Which is the best adaptive treatment strategy between two regimes that start with a different initial treatment?
4. Which is the overall best adaptive treatment strategy?

Research questions 1-3 are the most common for sizing a SMART, whereas the estimation of the overall embedded treatment strategy is mostly addressed as a secondary analysis in a more hypothesis generating fashion [Almirall et al., 2014]. Oetting et al. [2011] laid out the standard frequentist sample size formulae for continuous outcomes for each of these research questions.

In the next section, we examine these sample size derivations with particular focus on the third research question for continuous outcomes, as it is the topic of the first manuscript. If the final outcome of a SMART is binary, the Normal approximation can be used. Kidwell et al. [2018] derived ad-hoc sample size calculations for the third research question with binary outcomes. Several methodologies have also been developed to accommodate different types of outcome, such as longitudinal [Seewald et al., 2020, Dziak et al., 2021], cluster-level [NeCamp et al., 2017], and survival outcomes [Li and Murphy, 2011].

2.2.1 Research questions 1-3

The first and second research questions do not entail the comparison of treatment strategies. In particular, the sample size formula for the first research question is identical to that of a standard two-arm RCT where patients are pooled according to the first intervention. Taking the design on the left of Figure 2.1 as an example, the first research question entails the contrast in final outcomes of the patients who received treatment sequences 1-3 against those who received sequences 4-6. Let us indicate with A_j the treatment assigned at stage j and with Y the final outcome. In a one-sided hypothesis test with power $1 - \beta$, type I error α, and 1:1 randomization to the available treatments, the sample size required to test whether treatment a is a superior initial intervention than b is

$$n_1 = \frac{2(z_{1-\beta} - z_\alpha)^2}{\delta^2}, \qquad \delta = \frac{E(Y|A_1 = a) - E(Y|A_1 = b)}{\sqrt{\frac{Var(Y|A_1=a)+Var(Y|A_1=b)}{2}}}$$

where z_α indicates the quantile of the standard Normal distribution of level α. On the other hand, the second research question is of greater interest if the available second stage treatments for non-responders are the same, i.e., $e = c$ and $f = d$ in the left design depicted in Figure 2.1. Indicating with R the response indicator to the first treatment (1 for responders and 0 for non-responders), and assuming that the probability of response p to the first stage interventions is the same, that is, $p = \Pr(R = 1|A_1 = a) = \Pr(R = 1|A_1 = b)$, the sample size requirement for a one-sided test is

$$n_2 = \frac{2(z_{1-\beta} - z_\alpha)^2}{(1-p)\delta^2}, \qquad \delta = \frac{E(Y|R = 0, A_2 = c) - E(Y|R = 0, A_2 = d)}{\sqrt{\frac{Var(Y|R=0,A_2=c)+Var(Y|R=0,A_2=d)}{2}}}.$$

The sample size above is equivalent to that of a standard RCT where the patients' outcomes are pooled according to the appropriate sequence of treatments received (2 and 5 against 3 and 6 in Figure 2.1) and scaled by the probability of non-response to the first intervention.

Likewise, the sample size formula for the third research question can be reduced to a similar expression. To compare the ATS 'assign a and, if there is no response, switch to c' against 'assign b and, if there is no response, switch to e', the required sample size is

$$n_3 = \frac{2(z_{1-\beta} - z_\alpha)^2}{\delta^2} 4[2(1-p) + p], \qquad \delta = \frac{E(Y|A_1 = a, A_2 = c) - E(Y|A_1 = b, A_2 = e)}{\sqrt{\frac{Var(Y|A_1=a,A_2=c)+Var(Y|A_1=b,A_2=e)}{2}}}.$$

However, in addition to assuming the balanced allocation to the available treatments at each stage and the equal response rate to the initial treatments, this estimation relies on an assumption regarding the variance of the strategy outcomes that is difficult to test and prone to misspecification. More specifically, this assumption states that "the variability of the outcome around the strategy mean among either responders or non-responders is no more than the variance of the strategy mean". This assumption is expressed in mathematical terms in Equation 3.1. While the assumption of equal response rates can be avoided by either specifying the most conservative rate or fixing it to 1 (albeit generating a potentially overly conservative sample size), the variance assumption is necessary to achieve a formula that can be expressed in terms of the standardized mean difference (SMD). That is, the variance assumption allows the trialist to avoid specifying parameters that are likely poorly understood, such as the variance of the estimated strategy means. In the first manuscript, we build on this work to derive a methodology for sample size estimation to address the third research question that does not rely on either assumption. Both Oetting et al. [2011] and our approach rely on the estimation of the strategy mean via semiparametric marginal mean (MM) models [Murphy et al., 2001, Murphy, 2005b]. Dawson and Lavori [2010, 2012] argued that the MM variance estimator might be upwardly biased and lead to conservative sample size estimates. The authors derived an alternative sample size approach for SMARTs which relies on a different estimator based on maximum likelihood that they deemed more efficient. With respect to the derivations in Oetting et al. [2011], this approach has the advantages of not requiring the specification of the response rates to the initial treatments and allowing the comparison of strategies that start with the same initial treatment, but a similar assumption on the variance components is required, and this formulation also necessitates the specification of a variance inflation factor which might be difficult to elicit at the design stage of the trial. Another alternative approach was introduced in Ogbagaber et al. [2016]. The authors derived two sample size methodologies that allow for multiple pairwise comparisons between ATSs and the overall testing for the equality of all treatment strategies embedded in the SMART study. Their approach relies on the specification of the outcome means and variances of the treatment sequences embedded in the treatment strategies under investigation. The authors argue that while the number of unknown parameters to be specified is higher with respect to the methodology proposed by Dawson and Lavori [2010, 2012], information on the sequence of treatments is easier to obtain from observational data or non-SMART trials compared to the more complex variance inflation factor, and the elicitation of these quantities might be more reasonable than the assumption of equal response rates to the first stage interventions necessary for the calculations by Oetting et al. [2011]. However, it remains uncertain whether estimates from non-SMART trials can be confidently borrowed given the design differences that make SMARTs unique, and the authors also note that their methodology is more focused on multiple pairwise comparisons.

2.2.2 Research question 4 and other methodologies

The detection of the overall best embedded treatment strategy in a SMART is generally addressed as a secondary analysis. However, several methodologies have been introduced to size a SMART study for this task. Oetting et al. [2011] proposed a simulation-based algorithm that determines the sample size that allows the detection of the treatment strategy that results in the highest mean outcome with a desired probability. Ertefaie et al. [2016] adapted the 'multiple comparisons with the best' (MCB) methodology to determine the sample size required to estimate a set of ATSs that contains the best adaptive treatment strategy with a pre-specified probability. This work was further developed by Artman and co-authors [2020, 2022]. Rose et al. [2019] proposed two sample size approaches for the estimation of the optimal strategy with different degrees of modeling assumptions: one method relies on strong assumptions regarding the underlying data generating mechanism, whereas the second is based on bootstrap oversampling and relies on weaker assumptions and the availability of data from a SMART pilot.

Finally, the sample size methodologies discussed thus far concern the full-scale trial, but specific calculations are needed to size a pilot SMART. Typically, the primary aim of a pilot study is to evaluate the feasibility of conducting a full-scale trial. In a SMART study, given the sequential randomization and its dependence on intermediate responses, the number of subjects observed in each sequence of treatments is a random variable. To assess the feasibility of the trial, it is crucial to observe enough subjects in each subgroup. Hence the most common approach to sizing a pilot SMART is to estimate the minimum number of participants which ensures that a sufficient number of subjects is observed in each treatment sequence with a pre-specified probability [Almirall et al., 2012, Kim, 2016]. A precision-based alternative approach has been proposed by Yan et al. [2021], where the sample size is selected with the aim of estimating the marginal mean outcome of a treatment strategy within a pre-specified margin of error.

2.3 Bayesian sample size determination

Given the pre-experimental nature of the problem of determining the sample size of a clinical study, it is easy to see why a Bayesian approach can be attractive to statisticians. Even in a frequentist setting, one must make sharp assumptions regarding the effect of treatment based on current knowledge – i.e., a specific, single value must be posited, maybe derived from previous studies, whilst being in a setting where prior knowledge may be limited allowing for the equipoise that is essential for the ethical conduct of a trial. The Bayesian framework offers the tools and a streamlined course of action to incorporate such knowledge as well as the degree of confidence that we attribute to it through the elicitation of prior distributions.

2.3.1 A brief overview

The literature on Bayesian sample size determination in randomized controlled trials is vast, but it can be classified according to two schools of thought depending on whether the experiment is seen as a decision or an inference problem [Adcock, 1997, Spiegelhalter et al., 2004]. The decision-theory approach requires the definition of a utility function that leverages several factors, e.g., the cost of conducting a study and the benefits of the treatment under investigation, and the sample size that maximizes the expected utility is selected. On the other hand, in inference-based methods, also called performance-based [Wang et al., 2002, Brutti et al., 2014] or proper Bayesian [Spiegelhalter et al., 2004], the sample size is determined based on the inferential performance of a functional of the posterior distribution of the parameter of interest. In practice, the decision-theory approach has received less attention, as the definition of a utility function is complex and requires assumptions that make the methodology unrealistic for some types of trials [Spiegelhalter et al., 2004]. While it can be argued that the performance-based methods can be derived as special cases of decision-theory methods, the debate between the two schools of thought is fiery and outside of the scope of this thesis. For more information on the differences between approaches, see Lindley [1997], Joseph and Wolfson [1997], Adcock [1997], Pezeshk [2003].

Several performance-based approaches have been developed which differ in regard to the choice of the functional of the posterior distribution of the parameter of interest. For example, the average coverage criterion [Adcock, 1988], average length criterion [Joseph and Belisle, 1997], and worst outcome criterion [Joseph and Belisle, 1997] seek to control the properties of the credible interval of the parameter of interest in terms of coverage and/or length. The average posterior variance criterion aims to select the sample size which limits the posterior variance within an upper bound with a pre-specified probability Wang et al. [2002]. For a more detailed review of these and other methods, see Pezeshk [2003], Cao et al. [2009], Brutti et al. [2014]. Our interest is in the power criteria, which is the only method among those mentioned thus far with a frequentist counterpart.

Let us indicate with θ the parameter of interest of the study, with $\pi_0(\theta)$ its prior distribution (*analysis prior*), and consider the system of hypotheses

$$\begin{cases} \mathrm{H}_0 : \theta \in \Theta_0, \\ \mathrm{H}_1 : \theta \in \Theta_1. \end{cases}$$

We utilize the definition of Bayesian significance and power given in Spiegelhalter et al. [2004]. A result is considered significant if the posterior probability that θ belongs to the alternative hypothesis space Θ_1 is not inferior to a pre-specified threshold $1-\epsilon$, i.e.,

$$\Pr_{\pi(\cdot|\mathrm{Data})}(\theta \in \Theta_1) \geq 1-\epsilon, \qquad \epsilon \in (0,1).$$

The Bayesian power function is defined as the probability of obtaining a significant result under the probability measure associated with the sampling distribution of the random result of the experiment conditioning on a design value $\theta_d \in \Theta_1$, that is,

$$\eta(n) = \Pr_{f(\cdot|\theta_d)} \left\{ \Pr_{\pi(\cdot|\mathrm{Data})}(\theta \in \Theta_1) \geq 1-\epsilon \right\}. \tag{2.1}$$

2.3.2 The ‘two priors’ approach

In order to compute the Bayesian power function 2.1, one should make a sharp assumption on the value of the design parameter θ_d under the alternative hypothesis, and the sample size is selected as the minimum value of n which results in a power level greater than a specified threshold. This is equivalent to selecting a treatment effect or standardized mean difference under the alternative hypothesis in frequentist sample size determination. However, this approach leads the power analysis to depend critically on a single value, an issue known in Bayesian literature as local optimality [Brutti et al., 2008]. The Bayesian framework offers a straightforward path to alleviate this dependence on the availability of unrealistically precise information via the elicitation of a second prior distribution $\pi_d(\theta)$ on the design parameter, called the *design prior*. Note that the analysis prior π_0 and the design prior π_d have different roles and do not need to coincide. The analysis prior is the ‘classical’ prior distribution used in Bayesian inference for the analysis stage. It can incorporate pre-trial knowledge or it can be non-informative. Conversely, the design prior formalizes the uncertainty around the design parameter that is used to power the study. It follows that π_d must be informative and most of its mass has to lie in the alternative hypothesis space. By averaging the conditional sampling distribution of the random result V_n of the experiment over the design prior, the Bayesian power function becomes

$$\eta(n) = \Pr_{m_d(\cdot)} \left\{ \Pr_{\pi(\cdot|\text{Data})}(\theta \in \Theta_1) \geq 1 - \epsilon \right\},$$

$$m_d(v_n) = \int_{\Theta} f(v_n|\theta)\pi_d(\theta)d\theta.$$

This methodology is known as the 'two priors' approach. According to a review by Brutti et al. [2014], the idea of separate prior distributions for the design and analysis stages of an experiment was first postulated by Tsutakawa [1972] and further discussed and popularized by Wang et al. [2002]. This concept was then elaborated into the 'two priors' approach in following publications [Sahu and Smith, 2006, De Santis, 2006, Brutti et al., 2014, Sambucini, 2017].

Note that if the analysis prior is non-informative and a point-mass distribution is chosen for the design prior, the Bayesian and frequentist power functions coincide. For a more in-depth discussion of the relationship between frequentist and Bayesian sample size determination, see Inoue et al. [2005]. Additionally, even if a Bayesian approach is used to determine the sample size of a study, the final analysis stage can still be entirely frequentist. The class of approaches that employ this strategy is called 'hybrid frequentist-Bayesian'. For a review of these methods, see Kunzmann et al. [2021].

2.4 Recruitment forecasting in multicenter studies

Forecasting recruitments is a fundamental tool in the planning and monitoring stages of multicenter clinical studies, as it drives decisions with significant practical and financial consequences. Although the number of clinical studies and their associated costs have been

rising year on year for at least the past two decades [Martin et al., 2017, Gresham et al., 2020], the successful recruitment of the target number of participants within the planned timeline has remained a long-standing issue [Bieganek et al., 2022]. A considerable number of studies fails to reach the target sample size, which generally leads to the extension of the recruitment period, revising the sample size, the opening of new recruitment centers, or the discontinuation of the study. After interviewing investigators from 78 primary care research studies conducted in the Netherlands between 1999 and 2003, van der Wouden et al. [2007] noted that more than half (51%) failed to recruit the targeted sample size within the planned timeline and had to extend the fieldwork period. Similar results were found by Walters et al. [2017] and Jacques et al. [2022] in their respective reviews of publicly funded randomized clinical trials conducted in the United Kingdom. The former analyzed 151 randomized clinical trials conducted between 2004 and 2016, concluding that 66 (44%) did not achieve the original sample size, while the latter reached a similar number (47%) after reviewing 388 RCTs conducted between 1997 and 2020.
While reducing the sample size results in a decrease of power and extending the recruitment period may lead to a significant financial loss, discontinuing a clinical study also raises ethical concerns over the waste of often scarce research resources, especially if the results remain unreported. Kasenda et al. [2014] reviewed 1017 RCTs conducted in Switzerland, Germany, and Canada focusing on trial discontinuation and its causes. The authors found that study discontinuation due to reasons other than early apparent benefit is one of the main factors for the non-publication of the results, with poor recruitment being cited as the leading cause of trial discontinuation.

Yet, despite a growing body of literature on this topic, deterministic approaches based on the study investigators' expectations are still widely used in practice. In addition to ignoring significant sources of variability that a slightly more rigorous statistical model would address, the investigators' forecastings on recruitment rates are known to be overly optimistic. On one hand, investigators tend to overestimate the number of participants who meet the inclusion criteria and are willing to enroll in the study. This phenomenon takes the name of the Lasagna Law [Lasagna, 1979] and, in the 40 years since its postulation, it has mostly been shown to hold [van der Wouden et al., 2007, Bogin, 2022]. On the other hand, external pressure from stakeholders can lead the investigator to knowingly suggest higher than realistic recruitment rates in the planning phases in order to make the trial more appealing to the funder. Gkioni et al. [2020] recently surveyed chief investigators and statisticians involved in the planning and monitoring phase of the recruitment stage of studies conducted in Europe and the UK. Of the 23 investigators who responded, nine admitted that the pressure of making the study more attractive to the funder had an impact on the recruitment rates that were used for predicting enrollments. Furthermore, a staggering 90% of the surveyed statisticians (62/69) stated that they did not use any statistical model to forecast recruitment, and 41% were not aware of any statistical approach to predicting enrollments. Among statisticians who declared not using a statistical approach, the simplicity of a deterministic model, the non-familiarity with either the available statistical methods or their implementation, and doubts over the additional value of these methods were cited as the main reasons for not using them.

2.4.1 The Poisson and Poisson-Gamma models

Lee [1983] was the first to introduce a formal statistical context for the forecasting of recruitments. Indicating with λ the recruitment rate per time unit, the author proposed to model the participant arrival as a Poisson random variable. Hence, indicating with $N(t_j, t_i)$ the number of participants enrolled between time t_j and t_i, it follows that

$$\begin{aligned} &N(t_j, t_i) \sim \text{Poisson}[\lambda(t_i - t_j)], \\ &E[N(t_j, t_i)] = \lambda(t_i - t_j). \end{aligned}$$

The author focused on the estimation of interim recruitment goals between the start and end of the enrollment period and the probability of achieving such goals at each planned interim time. The average recruitment rate λ is assumed to be constant over time. In the planning phase of the study, an expected rate has to be elicited, whereas at the interim times it is estimated based on the observed enrollments, and the Normal approximation to the Poisson distribution is used to make predictions. Soon after, Moussa [1984] laid out a computer algorithm to implement this method. Senn [1998] and Carter [2004] further discussed the use of a Poisson process to model recruitments, shifting the focus on the estimation of the time required to reach the planned sample size and the probability of achieving this target.

A significant limitation of these methods is the underlying assumption that the centers recruit participants at the same fixed rate, which is typically not the case in practice. To solve this issue, Anisimov and Fedorov [2007a, 2007b] proposed to add a second layer of variability by viewing the recruitment rates as a sample from a Gamma distribution, introducing the doubly-stochastic Poisson-Gamma (PG) recruitment model. Indicating with C the number of centers and with u_i the initiation date of center i, the participants' arrival process to each center at time t is assumed to follow a Poisson process with random rate $\lambda_i(t) = I_{\{t>u_i\}}\lambda_i$, where I represents the indicator function and the rates λ_i's are considered an independent sample from a Gamma distribution with shape α and rate β. Therefore, the number of participants recruited up to time t in center i, i.e., $N_i(t)$, is a Poisson distributed with cumulative rate $\Lambda_i(t) = I_{\{t>u_i\}}(t - u_i)\lambda_i$. Given the independence between centers, the total number of recruitments $N(t) = \sum_C N_i(t)$ is still Poisson distributed. Hence, the PG model can be represented as

$$
\begin{aligned}
&N(t)|\Lambda(t) = \text{Poisson}[\Lambda(t)],\\
&\Lambda(t) = \sum_{i=1}^{C} \Lambda_i(t) = \sum_{i=1}^{C} I_{\{t>u_i\}}(t - u_i)\lambda_i,\\
&\lambda_i \sim \text{Gamma}(\alpha, \beta).
\end{aligned}
$$

This is an empirical Bayes method. Once an interim time t_{int} is reached, the accrued recruitment data collected in $[0, t_{int}]$ is used to estimate the hyperparameters of the prior Gamma distribution. This can be achieved by marginalizing the conditional distribution of $N_i(t)$, which leads to a Negative Binomial distribution, and the MLE estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the hyperparameters are obtained from the resulting likelihood. These estimates are then plugged into the posterior distribution of the cumulative rate, which can be represented as

$$\widetilde{\Lambda} = \sum_{i=1}^{C} \text{Gamma}(\widehat{\alpha} + k_i,\, \widehat{\beta} + t_{int} - u_i),$$

where k_i is the observed number of recruitments in center i up to the interim time. The authors show that if the recruitment centers share the same initiation date, then the remaining time to reach the targeted sample size is distributed as a Pearson type VI distribution. However, if the initiation dates are staggered, a closed form for this distribution is not achievable and Monte Carlo simulations are needed. Alternatively, one could compute the marginal expected value and variance of the additional number of participants recruited after the interim time $N^a(t)$ and achieve a point estimate and its associated credible interval via the Normal approximation for a grid of future time points. The credible intervals can then be inverted to achieve prediction intervals for the remaining time to complete the enrollment phase. This is quite straightforward, as the expected value and variance of future additional enrollments at a later point in time T are easily calculated. Assuming, for simplicity, that all the centers have started enrolling by the interim time, these quantities take the following form:

$$\begin{aligned} E[N^a(T)] &= (T - t_{int}) \sum_{i=1}^{C} \frac{\alpha + k_i}{\beta + t_{int} - u_i}, \\ Var[N^a(T)] &= (T - t_{int}) \sum_{i=1}^{C} \frac{\alpha + k_i}{\beta + t_{int} - u_i} + (T - t_{int})^2 \sum_{i=1}^{C} \frac{\alpha + k_i}{(\beta + t_{int} - u_i)^2}. \end{aligned}$$

This method has been validated on data from several real clinical trials [Anisimov and

Fedorov [2005], Anisimov and Fedorov [2007a], Anisimov [2009b], Zhang and Huang [2022]] and expanded in numerous directions. Anisimov and co-authors augmented the model to analyze the effects of the opening of new centers in an adaptive technique [Anisimov and Fedorov, 2007a] and their closure [Anisimov, 2011a], or to account for recruitment pauses in clinical trials with waiting time to response [Anisimov, 2011b]. In Anisimov [2011a], the author also introduces metrics to evaluate the performance of specific centers or regions and the effects of center-stratified randomization, which is further discussed in Anisimov [2011c]. Mijoule et al. [2012] investigated the use of the Pareto distribution in place of the Gamma distribution in a Pareto-Poisson model, concluding their study in favor of the Poisson-Gamma model due to the small differences between models and the simplicity of the PG method. The authors also conducted a sensitivity analysis to assess the impact of errors in the estimation of hyperparameters and discussed the use of a Uniform distribution to model unknown initiation times, which had already been briefly discussed in Anisimov [2009a]. Minois et al. [2017b] modified the PG model to allow for breaks in the recruitment process such as weekends or holidays using a piecewise constant rate, however the authors concluded that the standard PG model still yielded better results. Anisimov et al. [2022] augmented the PG approach to include several methods of increasing complexity to model participants' drop-out either upon arrival or during the screening period.

Finally, it is important to mention the work of Gajewski et al. [2008], where the authors proposed a Bayesian method where the waiting times between the participants' arrival are modeled as an exponential random variable and an informative Inverse Gamma distribution is placed on its parameter. The authors focused on the planning phase rather than the monitoring phase of the enrollment process, hence it is necessary to incorporate subjective knowledge from experts to elicit the parameters of the prior distribution. While this methodology does not directly model the participants' arrival as a Poisson process, assuming an exponential distribution for the waiting times is equivalent. However, this model is only applicable for trials with one center.

Before moving on to the next section on time-dependent models, it is important to clarify what is meant in this thesis by time dependence. Some of the authors referenced thus far refer to their method as time-dependent, or to the Poisson process as non-homogeneous. This is due to centers having staggered initiation times, as sites that have a delayed starting date will have a recruitment rate of 0 until their openings. However, once they open, the rates are assumed to be constant over time. In practice, this may be unlikely, particularly when studies are recruiting from relatively small or fixed populations (e.g., eligibility criteria include a diagnosis of a moderately rare disease) such that recruiting sites may exhaust the pool of potential participants over time.

This assumption of a constant rate of recruitment once a center has opened is what the methods that are presented in the next section seek to weaken. Throughout the rest of this thesis, methodologies that do not rely on the constant-rate assumption are referred to as time-dependent.

2.4.2 Time-dependent models

The constant-rate assumption is often unrealistic in practice, and several methods have been introduced in recent years to alleviate this concern. Tang et al. [2012] proposed a discrete-time Poisson process based approach whereby the overall recruitment rate varies with time until an unknown time point estimated via a changepoint analysis, after which it stabilizes, and then it increases towards the last phase of recruitment. This method mainly focuses on the very last period of the recruitment phase, and further assumptions are needed on the level of increased recruitments and the future point in time where such an increase will be observed.

A more flexible model was introduced in Zhang and Long [2010]. The authors proposed a Bayesian model where the overall trial accrual is captured via a non-homogeneous Poisson process and the underlying time-dependent recruitment rate is modeled through cubic B-splines. Once an interim time is reached, multiple models are fitted to the data varying the number of equally spaced knots of the cubic B-spline basis functions, and the best-fitting model to be used for predictions is selected via the deviance information criterion. A posterior distribution for the overall recruitment rate of the study at the interim time is estimated and the prediction on future enrollments is carried out assuming that the rate will be constant beyond that point. This methodology was also extended to allow for the prediction of event times in Zhang and Long [2012]. Although this method is flexible in that it allows for the non-parametric estimation of the overall recruitment rate, there is no guarantee that the rate will be constant beyond the monitoring time. While the stabilization of the recruitment intensity at some point is a common assumption, if predictions are carried out before such a point is passed, they may be biased. The authors argue that since the recruitment intensity generally increases with time at the early stages, a biased estimate would typically lead to underestimating recruitments, which is a more acceptable risk compared to an overestimation. Additionally, as the model proposed in Tang et al. [2012], this approach does not account for heterogeneity between centers and does not directly take into account staggered initiation times, both of which can have a substantial impact on the recruitment process, particularly if the study is conducted in different regions/countries. Deng et al. [2017] alleviated this concern by stratifying the estimation of the underlying recruitment rate by region.

A time-dependent model that relies on the Poisson-Gamma setting was introduced in Lan et al. [2019]. The authors proposed a simulation-based approach where the recruitment intensity is assumed to be constant in the initial stage following the center's opening and then it decays over time as a negative exponential. The authors augmented the model by including a center initiation model to account for future unknown openings of new centers and the optional inclusion of a final surge in enrollments similarly to Tang et al. [2012]. Urbas et al. [2022] further expanded this methodology by allowing for the fitting of different parametric curves to capture the monotonic decay in the recruitment rates which are then used for predictions via Bayesian model averaging. The authors also included a test for the detection of time-inhomogeneity in the underlying recruitment intensity.

2.4.3 Other methods

In Sections 2.4.1 and 2.4.2, we primarily focused on methods based on the Poisson process, as they currently are the most popular and most frequently implemented by a large margin. We did not discuss early deterministic models or methods based on Brownian motion, such as those proposed in Lai et al. [2001] and Zhang and Lai [2011]. For a more schematic and detailed comparison between some of the methods discussed so far and more, we refer the reader to the systematic reviews by Barnard et al. [2010], Heitjan et al. [2015], and Gkioni et al. [2019].

Furthermore, most of the methods presented in this section focus on the monitoring phase of the recruitment stage, as they require data from the ongoing recruitment process to produce forecastings. Although some of these methods can be adapted for use in the planning phase of the study, they require further assumptions and/or modifications. For example, Zhang and Long [2010] describe how their methodology can be adapted to be used before the start of the enrollment period, however, doing so would require the assumption of a constant recruitment rate and the elicitation of an informative prior distribution which captures the confidence regarding the assumed rate, essentially making their time-dependent methodology not too different from the use of a homogeneous Poisson process or the model from Gajewski et al. [2008].

Anisimov [2008] suggested how the Poisson-Gamma model can be used in the planning stage using the expected number of recruitments per region or historical data from similar studies to tune the hyperparameters α and β. Bakhshi et al. [2013] and Minois et al. [2017a] formalized this idea and developed two methods to incorporate prior information from concluded trials. The first proposed the addition of a third level of hierarchy in the Poisson-Gamma model to capture the between-trial variation. The authors outlined a random-effects meta-analysis based approach whereby the orthogonally reparametrized parameters of the Gamma distribution are estimated using data from a number of concluded trials.

However, the authors concluded that while this method could be feasible, they were not able to implement it in practice due to the excessive variability between centers, which raises concerns over the unrealistic underlying exchangeability assumption between trials. For this reason, Minois et al. [2017a] focused on the selection of only one historical trial that is as close as possible to the new trial in terms of therapeutic area, inclusion/exclusion criteria, regions, and participants' characteristics. If enrollment data on such a trial are available, then the hyperparameters of the Gamma distributions can be estimated from the centers that are shared between the two studies and used to predict enrollments for these sites, while additional assumptions are needed to forecast recruitments in the unshared centers.
For a more in-depth description of models that can be used in the planning phase of a study, see the systematic review by Gkioni et al. [2019].

2.5 Summary

In this chapter, we have reviewed the literature that contextualizes the three manuscripts of this thesis. We have discussed the advantages of the SMART design compared to other experimental designs and its challenges, especially in regard to sample size determination. We have discussed the existing methodologies for sizing SMARTs and their assumptions, with particular focus on the comparison between strategies that start with a different intervention. We have given a brief overview of the Bayesian literature on sample size determination and highlighted how the Bayesian framework allows for the incorporation of uncertainty around unknown parameters at the pre-experimental stage. Finally, we have discussed the importance of developing and implementing statistical methodologies for forecasting recruitments in clinical studies and reviewed the existing approaches.

Chapter 3

Bayesian sample size calculations for comparing two strategies in SMART studies

Preamble to Manuscript 1. The sample size determination for comparing two adaptive treatment strategies in a SMART study has mainly been analyzed in the frequentist setting. Oetting et al. [2011] outlined the standard frequentist calculations for continuous outcomes based on the semiparametric marginal mean model estimator of the strategy mean [Murphy, 2005b]. However, although the resulting formulae are similar to those of traditional RCTs, they rely on additional assumptions and specifications.

This chapter aims to extend these calculations to the Bayesian framework in order to achieve a more robust and flexible approach that relies on fewer assumptions. This chapter's contribution to original methodology consists of adapting the Bayesian 'two priors' approach [Wang et al., 2002, Sahu and Smith, 2006] to the SMART design. To obtain a methodology that does not depend on the assumptions on the variance components and response rates to the first intervention posited in the frequentist approach, the Bayesian power function is marginalized over the posterior distribution of the variance components estimated on pilot data. This manuscript provides a simulation study where the proposed methodology is compared to the standard frequentist formula in different scenarios for varying levels of model misspecification. The applicability of the Bayesian approach developed in this chapter is demonstrated by sizing a SMART study that seeks to evaluate the efficacy of internet-based strategies for stress management using data from its pilot.

The manuscript presented in this chapter was published in *Biometrics* [Turchetta et al., 2022]. In addition to the supplementary material published with this manuscript presented in Appendix A, an online tutorial on the `bayesSMARTsize` R package developed to implement the proposed methodology is presented in Appendix B.

Bayesian sample size calculations for comparing two strategies in SMART studies

Armando Turchetta[1], Erica E.M. Moodie[1], David A. Stephens[2], Sylvie D. Lambert[3].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2]*Department of Mathematics and Statistics, McGill University*

[3]*Ingram School of Nursing, McGill University*

This thesis contains the accepted version of the corresponding paper published in *Biometrics* [Turchetta et al., 2022].

Abstract

In the management of most chronic conditions characterized by the lack of universally effective treatments, adaptive treatment strategies (ATSs) have grown in popularity as they offer a more individualized approach. As a result, sequential multiple assignment randomized trials (SMARTs) have gained attention as the most suitable clinical trial design to formalize the study of these strategies. While the number of SMARTs has increased in recent years, sample size and design considerations have generally been carried out in frequentist settings. However, standard frequentist formulae require assumptions on interim response rates and variance components. Misspecifying these can lead to incorrect sample size calculations and correspondingly inadequate levels of power. The Bayesian framework offers a straightforward path to alleviate some of these concerns. In this paper, we provide calculations in a Bayesian setting to allow more realistic and robust estimates that account for uncertainty in inputs through the 'two priors' approach. Additionally, compared to the standard frequentist formulae, this methodology allows us to rely on fewer assumptions, integrate pre-trial knowledge, and switch the focus from the standardized effect size to the minimal detectable difference. The proposed methodology is evaluated in a thorough simulation study and is implemented to estimate the sample size for a full-scale SMART of an Internet-Based Adaptive Stress Management intervention on cardiovascular disease patients using data from its pilot study conducted in two Canadian provinces.

3.1 Introduction

Precision medicine has become a popular topic in the field of healthcare [Kosorok and Moodie, 2015]. Within this medical model, treatments and decision rules are personalized and tailored to patient characteristics, shifting the focus from the traditional treatment of the diagnosis to the treatment of the patient [Wallace and Moodie, 2014]. In settings where there is a lack of a universally effective treatment, several interventions are often needed to prevent onset and alleviate symptoms improving the patient quality of life, requiring a sequential, individualized approach whereby interventions are adapted and re-adapted over time in response to the specific needs and evolving condition of the individual [Kosorok and Laber, 2019].

In order to estimate the optimal individualized sequence of treatments for each patient, adaptive treatment strategies (ATSs), also known as dynamic treatment regimes (DTRs), have been introduced [Lavori et al., 2000, Murphy et al., 2001]. To formalize the study of these regimes, the sequential multiple assignment randomized trial (SMART) has been developed [Lavori and Dawson, 2000, Murphy, 2005b]. SMARTs are based on multiple stages, each representing a clinical decision point: at each step, the patients are randomized accounting for a small set of characteristics or responses to previous interventions. SMARTs are considered the gold standard for developing adaptive treatment strategies, however, because of the cost necessary to fund an adequately powered SMART, observational studies are still largely employed to develop ATSs, even though the use of such data entails an additional layer of complexity to mitigate confounding [Chakraborty and Moodie, 2013]. Because of their nature, different types of SMARTs can be defined by varying the number of stages and available treatments at each stage [Lei et al., 2012]. We focus on the two-stage design outlined in Figure 3.1 where responders to the first stage intervention continue with the same treatment whereas non-responders are re-randomized. Note that the term ‘adaptive’ pertains to the treatment allocation of the strategies being analyzed and not the trial design, which is fixed and does not entail mid-trial adjustments of the randomization probabilities or other design parameters, nor does it involve interim analyses of the data.

When compared to standard randomized controlled trials (RCTs), SMARTs have some relevant advantages: most importantly, they allow for the direct comparison of multiple ATSs and to discover interactions between treatments. In this type of design, it is crucial to identify carry-over effects of previous treatments on the future ones, so as to avoid the detection of interventions that only appear to be optimal in the short term but are not in fact optimal in the long term, e.g. because they may preclude later, more effective therapies. SMARTs are designed to identify interactions that regular RCTs are likely to miss, as the latter are typically powered to make comparisons between average effects in each treatment arm. From an operational point of view, SMARTs share some features with crossover trials, as in both designs subjects are assigned to a sequence of treatments. However, they are conceptually different, in that the latter design is characterized by a washout period between treatments to allow for any carry-over effect to wear off, removing the possibility to detect interactions. Additionally, in crossover designs, treatments are not allocated according to intermediate responses to previous interventions. This tailoring feature of SMARTs is also what distinguishes them from other simpler designs that are operationally similar, such as multi-arm trials where patients are randomized to fixed sequences of treatments or factorial trials with a delayed second randomization. Furthermore, due to their randomization process and the variety of treatments, SMARTs are ethically advantageous and appealing to the study participants [Wallace et al., 2016]. So far, SMARTs have been deployed to estimate optimal strategies in a wide range of fields, such as weight loss [Almirall et al., 2014], substance abuse [Murphy et al., 2007], and cancer [Kidwell, 2014, Sikorskii et al., 2017], with particular emphasis on prostate cancer [Wang et al., 2012]. Notably, because of the adaptive nature of the treatments under consideration, SMARTs have assumed an important role in the management of chronic diseases [Chakraborty and Moodie, 2013], namely ADHD [Pelham Jr et al., 2016], schizophrenia [Stroup et al., 2003], and alcohol dependence [Nahum-Shani et al., 2017], among others. In this paper, for example, we will apply our proposed methodology to estimate the sample size of a web-based stress management program study, named Internet-Based Adaptive Stress Management SMART [Lambert et al., 2021], which employs the SMART design to overcome the necessity for interventions that are tailored to the patients' needs, which often lead to better outcomes in internet-based programs. While the number of SMARTs has increased in recent years, and it is clear they have the potential for playing an important role in the management of chronic conditions, their theoretical features are yet to be fully discovered. One of the key elements to be established in the design phase of a SMART is the definition of the research question. In fact, due to its structure, there are several research questions that can be addressed, the most common of which are (1) Which is the best initial treatment? (2) Which is the best secondary treatment among non-responders to the first intervention? (3) Which is the best ATS between two strategies that start with a different initial treatment? (4) Which is the overall best ATS? Several sample size estimation methods for determining the overall optimal ATS have recently been introduced [Oetting et al. [2011], Rose et al. [2019], Artman et al. [2020], Artman et al. [2022]], however, given the additional complexity induced by the multiple comparisons between all the embedded regimes in a SMART, the comparison of more than two strategies is currently mainly performed as a secondary analysis or in exploratory analyses aimed at generating hypotheses to be assessed in subsequent confirmatory trials. Most of the primary analyses performed on SMARTs are focused on the comparison between two means or two strategies, and given that the mean outcome of a strategy is a weighted mean across outcomes of individuals whose paths are consistent with the strategy, frequentist calculations for SMARTs sample sizes with continuous outcomes are similar to traditional randomized clinical trials [Oetting et al., 2011, Kosorok and Moodie, 2015, Kidwell et al., 2018]. In this paper, we focus on the third research question, as it is the most common type of analysis addressed in a SMART that entails the comparison of two strategies. Despite their similarity with more classical RCTs, SMART sample size calculations generally rely on additional assumptions and specifications of key design parameters. Yet, little attention has been paid to the robustness of these calculations to model misspecification or uncertainty. In particular, response rates to initial treatments and a standardized effect size must be fixed, leading to a decrease in power if responses are misspecified or if there is variability around them. In Bayesian literature, this shortcoming is commonly known as local optimality, and a series of hybrid frequentist–Bayesian and fully Bayesian approaches have been introduced to overcome this drawback [Spiegelhalter et al., 2004, Kunzmann et al., 2021]. One of the methods that belongs to the latter category is the 'two priors' approach, which can be seen as a flexible and useful extension of standard frequentist methods [Wang et al., 2002, Sahu and Smith, 2006, De Santis, 2006, Brutti et al., 2014, Sambucini, 2017]. In this paper, we adapt the 'two priors' approach to the SMART design, and we analyze the performance of this Bayesian method to sizing a SMART in a detailed simulation study. With respect to the standard frequentist formulae, the proposed method relies on fewer assumptions, allows for greater flexibility in the design stage, and leads to more robust sample size calculations. The major drawback of this approach lies in the Bayesian power function's dependence on the variance of the strategy means' estimator, which, contrary to frequentist calculations, needs to be specified. Since full-scale SMARTs are often preceded by a pilot study, the easiest course of action is to plug-in the estimates obtained from pilot data. However, the use of crude estimates from pilot studies of the variance components needed for sample size computations is controversial, as these studies are generally not sized to guarantee precise estimations of such parameters and might lead to underpowered full-scale trials [Browne, 1995, Vickers, 2003, Bell et al., 2018]. Specifically, SMART pilot studies are commonly sized to ensure that a sufficient number of subjects for each treatment sequence is observed with high probability [Kim, 2016] rather than through precision-based approaches (although an alternative has recently been introduced in Yan et al. [2021]). To overcome this pitfall, we propose to marginalize the Bayesian power function over the posterior distribution of the variance components estimated on pilot data in order to account for their variability. The paper is structured as follows. In Section 2, we give an overview of the frequentist sample size formula and we outline its Bayesian generalization. In Section 3, we analyze the performance of the proposed method in terms of power and type I error through an extensive simulation study. In Section 4, we apply our method using data from the Internet-Based Adaptive Stress Management Pilot SMART in order to estimate the sample size of its full-scale version. Section 5 concludes.

3.2 Methodology

Let us consider a SMART with k stages. We focus on the comparison of two adaptive treatment strategies that begin with different initial treatments. Let A_j be the treatment assigned at stage $j = 1, \ldots, k$, S_1 the pre-treatment information, and $\{S_j, j \geq 2\}$ the intermediate response indicator after treatment A_{j-1} (1 for responders and 0 for non-responders), so that the ordered data trajectory is $\{S_1, A_1, S_2, A_2, \ldots, S_k, A_k\}$. The overbar indicates the accrual of information up to the index, e.g. $\overline{S}_j = \{S_1, \ldots, S_j\}$. We denote with d_j the decision rule on treatment allocation at stage j. For each stage j, the decision rule takes as inputs the information collected up to that point, i.e. $\overline{S}_j$ and, if $j \geq 2$, the previous treatment allocations $\overline{A}_{j-1}$, and outputs the new treatment to be assigned A_j. An adaptive treatment strategy is a sequence of fixed decision rules that personalize the treatment sequence, and it is represented by $\bar{d}_k = \{d_1, \ldots, d_k\}$. The final continuous outcome recorded at the end of the trial is denoted by Y and n is the total sample size of the trial. If Y is binary, the Normal approximation can be used. Type I and type II errors are respectively indicated with α and β and z_α represents the quantile of the standard Normal distribution of level α.

In this paper, we consider the two-stage SMART design depicted in Figure 3.1. In this scheme, patients are first randomized to either treatment a or b. Afterward, responders to the first intervention will continue with the same treatment, whereas non-responders are re-randomized to their second stage treatment. Note that this design embeds six treatment sequences – $\{aa, ac, ad, bb, be, bf\}$ – and four different strategies of the form ‘assign treatment A_1 and, if the subject does not respond, switch to treatment A_2', where the candidate treatments are $\{a, b\}$ for the first stage and $\{c, d\}$ or $\{e, f\}$ for stage 2.

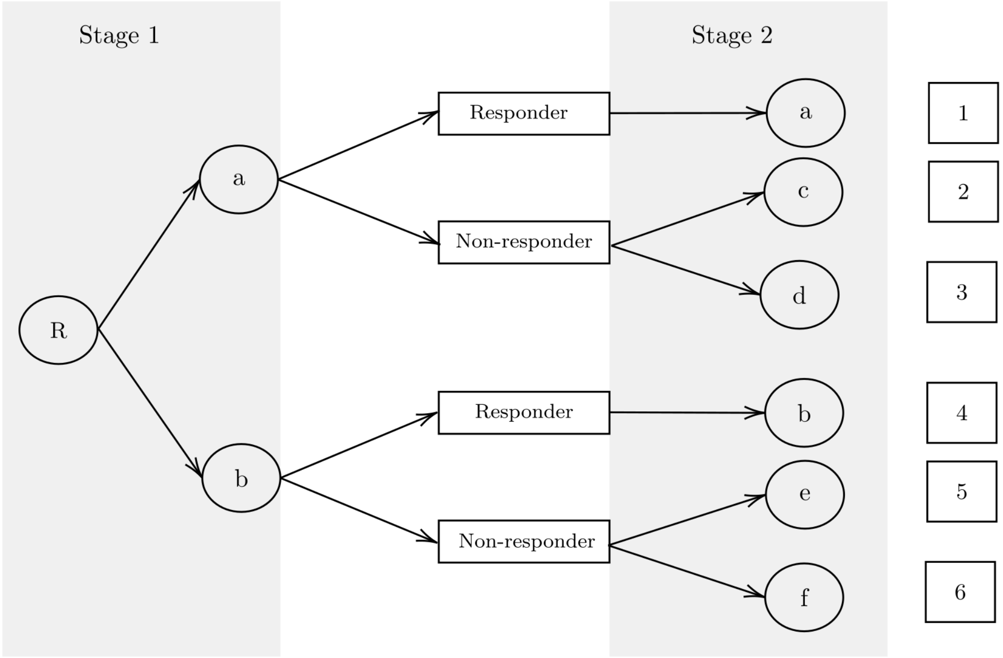

Figure 3.1: SMART scheme. This design embeds six treatment sequences and four ATSs of the form 'assign A_1 and, if there is no response, switch to A_2'.

3.2.1 Frequentist sample size estimate

Using results from Murphy et al. [2001] and Murphy [2005b], under the assumption that at any stage, and for any given history, the probability of any treatment included in an adaptive treatment strategy being assigned is positive, a consistent estimator of the mean outcome Y under strategy $\bar{d}_k$, which we denote $\mu_{\bar{d}_k}$, is

$$\widehat{\mu}_{\bar{d}_k} = \frac{\mathbf{P}_n\left[\prod_{j=1}^{k} \frac{I\left\{\overline{A}_j = d_j(\overline{S}_j, \overline{A}_{j-1})\right\}}{\Pr(d_j|\overline{S}_j, \overline{A}_{j-1})} Y\right]}{\mathbf{P}_n\left[\prod_{j=1}^{k} \frac{I\left\{\overline{A}_j = d_j(\overline{S}_j, \overline{A}_{j-1})\right\}}{\Pr(d_j|\overline{S}_j, \overline{A}_{j-1})}\right]}$$

where $\mathbf{P}_n$ represents the sample average and I the indicator function. Furthermore, defining

$$U(\overline{S}_k, \overline{A}_k, \bar{d}_k, \mu_{\bar{d}_k}) = \prod_{j=1}^{k} \frac{I\left\{\overline{A}_j = d_j(\overline{S}_j, \overline{A}_{j-1})\right\}}{\Pr(d_j|\overline{S}_j, \overline{A}_{j-1})}(Y - \mu_{\bar{d}_k}),$$

a consistent estimator of the variance of $\sqrt{n}(\widehat{\mu}_{\bar{d}_k} - \widehat{\mu}_{\overline{d'}_k})$ is $\mathbf{P}_n(U^2(\overline{S}_k, \overline{A}_k, \bar{d}_k, \mu_{\bar{d}_k}) + U^2(\overline{S}_k, \overline{A}_k, \overline{d'}_k, \mu_{\overline{d'}_k}))$ and the test statistic

$$Z = \frac{\sqrt{n}(\widehat{\mu}_{\bar{d}_k} - \widehat{\mu}_{\overline{d'}_k})}{\mathbf{P}_n(U^2(\overline{S}_k, \overline{A}_k, \bar{d}_k, \mu_{\bar{d}_k}) + U^2(\overline{S}_k, \overline{A}_k, \overline{d'}_k, \mu_{\overline{d'}_k}))}$$

is normally distributed for large samples. By writing the variance of $\sqrt{n}\widehat{\mu}_{\bar{d}_k}$ as

$$\tau^2_{\bar{d}_k} = E_{\bar{d}_k}\left[\prod_{j=1}^{k} \frac{(Y - \mu_{\bar{d}_k})^2}{\Pr(d_j|\overline{S}_j, \bar{d}_{j-1})}\right],$$

if there is no pre-treatment information S_1 and we consider a two-stage SMART ($k = 2$), indicating the intermediate response S_2 with R, it follows that

$$\begin{aligned}\tau^2_{\bar{d}_k} = E_{\bar{d}_k}&\left[\frac{(Y - \mu_{\bar{d}_k})^2}{\Pr(A_1 = a_1)\Pr(A_2 = a_2|A_1 = a_1, R = 1)}\right]\Pr(R = 1) + \\ &E_{\bar{d}_k}\left[\frac{(Y - \mu_{\bar{d}_k})^2}{\Pr(A_1 = a_1)\Pr(A_2 = a_2|A_1 = a_1, R = 0)}\right]\Pr(R = 0).\end{aligned}$$

If the variable used to define the response status is continuous, a threshold or condition is needed to dichotomize it so as to identify responders to the first intervention. Assuming that

1. the variability of the outcome Y around the strategy mean for both responders ($R = 1$) and non-responders ($R = 0$) is not greater than the variance of the strategy mean, i.e.

$$E_{\bar{d}_k}\left[(Y - \mu_{\bar{d}_k})^2|R\right] \leq E_{\bar{d}_k}\left[(Y - \mu_{\bar{d}_k})^2\right]; \tag{3.1}$$

2. the response rates to the initial treatments are equal; and

3. at each stage, patients are allocated equally to the available treatments,

and considering the SMART design outlined in Figure 3.1 where responders to the initial treatments have only one subsequent treatment option, an upper bound of $\tau^2_{\bar{d}_k}$ is $2\sigma^2_{\bar{d}_k}(2-p)$, where p is the common response rate to the initial treatments and $\sigma^2_{\bar{d}_k}$ is the marginal variance of the strategy. Considering the system of hypotheses

$$\begin{cases} \mathrm{H}_0 : \mu_{\bar{d}^1_k} - \mu_{\bar{d}^2_k} = 0 \\ \mathrm{H}_1 : \mu_{\bar{d}^1_k} - \mu_{\bar{d}^2_k} > 0 \end{cases} \tag{3.2}$$

where $\bar{d}^1_k$ and $\bar{d}^2_k$ are two strategies with a different initial treatment, and using the upper bound of $\tau^2_{\bar{d}_k}$, for a standardized effect size $\delta = \frac{\mu_{\bar{d}^1_k} - \mu_{\bar{d}^2_k}}{\sigma}$ where $\sigma = \sqrt{\left(\sigma^2_{\bar{d}^1_k} + \sigma^2_{\bar{d}^2_k}\right)/2}$, the sample size formula is given by the value of n which satisfies $\Pr(Z > z_{1-\alpha} | \mu_{\bar{d}^1_k} - \mu_{\bar{d}^2_k} = \delta\sigma) = 1 - \beta$, which is

$$n = \frac{(z_\beta + z_\alpha)^2}{\delta^2} 4[2(1-p) + p]. \tag{3.3}$$

If the response rates to the initial treatments are known to be different, the lowest response rate should be used. In case they are unknown, the common response rate can be set to 0. Both options can lead to overly conservative estimates. In the next sections, we will first apply the ‘two priors’ approach to the SMART design, showing how prior beliefs on the uncertainty of design parameters can be included into the calculations, and then outline how to incorporate estimates of the variance components from pilot studies while accounting for their variability, resulting in a methodology that does not rely on Assumptions 1 and 2.

3.2.2 Applying the ‘two priors’ approach to the SMART design

In accordance with the definition of Bayesian significance given by Spiegelhalter et al. [2004], a result is considered significant if the posterior probability that the parameter of interest θ belongs to the alternative hypothesis space Θ_1 is not less than a specified threshold $1-\epsilon$ where $\epsilon \in (0,1)$, i.e. $\Pr_{\pi(\cdot|\text{Data})}(\theta \in \Theta_1) \geq 1-\epsilon$. Setting $V_n = \widehat{\mu}_{\bar{d}_k^1} - \widehat{\mu}_{\bar{d}_k^2}$, $\theta = \mu_{\bar{d}_k^1} - \mu_{\bar{d}_k^2}$, and $\tau^2 = \tau^2_{\bar{d}_k^1} + \tau^2_{\bar{d}_k^2}$ to ease the notation, for large samples, $V_n|\theta \sim \mathcal{N}\left(\theta, \frac{\tau^2}{n}\right)$. If we consider the conjugate prior distribution for θ, $\pi_0(\theta) = \mathcal{N}(\theta; \theta_0, \sigma_0^2)$, called *analysis prior*, the posterior distribution of the parameter of interest is

$$\pi(\theta|V_n) = \mathcal{N}\left(\theta; \frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2},\ \left(\frac{1}{\sigma_0^2} + \frac{n}{\tau^2}\right)^{-1}\right),$$

and it follows that the outcome of the clinical trial is significant in the Bayesian sense if $\Pr_{\pi(\cdot|V_n)}(\theta > 0) \geq 1-\epsilon$, i.e. when

$$V_n \geq -\frac{z_\epsilon \tau \sqrt{\tau^2 + n\sigma_0}}{n\sigma_0} - \frac{\theta_0 \tau}{n\sigma_0^2}. \tag{3.4}$$

The steps necessary to reach this inequality are laid out in Web Appendix A. Since in the pre-experimental phase V_n has not been observed yet, the Bayesian power function is defined as the probability of obtaining a Bayesian significant result. To compute this probability, similarly to frequentist calculations, the standard approach consists of using the distribution of V_n conditional on a value θ_d under Θ_1. In order to overcome the local optimality issue – i.e. optimal performance only under specific values of the design parameters, with performance losses under alternative specifications – the ‘two priors’ approach entails the elicitation of a second prior distribution $\pi_d(\theta)$, called *design prior*, which formalizes the uncertainty around θ_d. Since our parameter of interest is $\theta = \mu_{\bar{d}_k^1} - \mu_{\bar{d}_k^2}$, π_d encapsulates the variability around the minimal detectable difference (MDD) of the two strategies being compared, i.e. the average difference between strategy means. Setting the conjugate design prior $\pi_d(\theta) = \mathcal{N}(\theta; \theta_d, \sigma_d^2)$

and indicating with Φ the cumulative distribution function of the standard Normal random variable, the marginal distribution of V_n is $m_d(V_n) = \mathcal{N}\left(V_n; \theta_d, \frac{\tau^2}{n} + \sigma_d^2\right)$, hence the Bayesian power function $\eta(n; \tau^2)$ is $\Pr_{m_d(\cdot)}\{\Pr_{\pi(\cdot|V_n)}(\theta > 0) \geq 1 - \epsilon\}$ and it can be expressed as

$$\eta(n; \tau^2) = \Phi\left[\frac{1}{\sqrt{\frac{\tau^2}{n} + \sigma_d^2}}\left(\frac{\theta_0 \tau^2}{n\sigma_0^2} + \theta_d + \frac{z_\epsilon \tau \sqrt{\tau^2 + n\sigma_0^2}}{n\sigma_0}\right)\right]. \tag{3.5}$$

It is important to emphasize that the two priors π_0 and π_d serve different purposes. The former is the density commonly used in Bayesian inference for the analysis stage. It can formalize pre-trial knowledge or be non-informative. On the other hand, the design prior is employed in a what-if spirit to describe the scenario in which the sample size is determined. As such, this distribution has to be proper and most of its mass must lie in the alternative hypothesis space. Note that by pre-trial knowledge we mean the information that can be borrowed from previous studies to influence the analysis stage of the clinical trial. In a broader sense, the term could also be applied to the frequentist setting where such knowledge is useful in the postulation of the effect size and the response rate to the first stage interventions. However, this is a different use of pre-trial information that pertains to the design phase rather than the analysis stage of a trial, and its Bayesian counterpart in the 'two priors' approach is the conjecture of the design prior π_d. Note that, if $\sigma_d^2 \to 0$ and $\sigma_0^2 \to \infty$, Equation (3.5) reduces to the frequentist power function which leads to the sample size Formula (3.3) when the upper bounds of the variance of the strategy means' estimator are used.

3.2.3 Accounting for variability around the variance components estimates

A drawback of the Bayesian power function consists in the specification of $\tau^2_{\bar{d}^1_k}$ and $\tau^2_{\bar{d}^2_k}$. In fact, contrary to the frequentist counterpart, replacing $\tau^2_{\bar{d}^1_k}$ and $\tau^2_{\bar{d}^2_k}$ with their upper bounds does not lead to a simplified formula which allows us to avoid the direct specification of the variance components by specifying a standardized effect size. To overcome this pitfall and properly size a full-scale SMART, we propose the integration of prior knowledge from its pilot study. However, the direct use of a plug-in estimate of the variance components from pilot studies has been generally criticized, as it often leads to underpowered trials [Vickers, 2003, Browne, 1995, Bell et al., 2018]. In order to account for the uncertainty around the estimates of $\tau^2_{\bar{d}^1_k}$ and $\tau^2_{\bar{d}^2_k}$, instead of using their crude estimates, we propose the use of their posterior distribution based on pilot data to marginalize the Bayesian power function given in Equation (3.5). We assume that the pilot SMART is conducted on a population with similar characteristics to its full-scale version. Indicating with the superscript p the quantities that are pertinent to the pilot study, from the previous section we have that $V^p_n|\theta, \tau^2 \sim \mathcal{N}\left(\theta, \frac{\tau^2}{n}\right)$. A possible choice of prior conjugate distribution for (θ, τ^2) is the Normal-inverse-chi-squared (NIX) density with parameters $\theta_p, \kappa_p, \sigma^2_p$ and ν_p, where θ_p and σ^2_p represent the prior values of θ and τ^2, while κ_p and ν_p set the strength of the prior specifications [Murphy, 2007]. This density has the form of the product between a Normal distribution and the PDF of a Noncentral chi-squared random variable, i.e. $\pi(\theta, \tau^2) = \mathcal{N}(\theta|\theta_p, \sigma^2_p)\chi^{-2}(\tau^2|\nu_p, \sigma^2_p)$. It follows that the marginal posterior distribution $\pi(\tau^2|V^p_n)$ is a Noncentral chi-squared density with parameters

$$\nu_n = \nu_p + n, \qquad \sigma^2_n = \frac{1}{\nu_n}\left[\sigma^2_p\nu_p + n\widehat{\tau^p}^2 + \frac{n\kappa_p}{\kappa_p + n}\left(\theta_p - \widehat{\theta}^p\right)^2\right],$$

where $\widehat{\theta}^p = \widehat{\mu}^p_{\bar{d}^1_k} - \widehat{\mu}^p_{\bar{d}^2_k}$ and $\widehat{\tau^p}^2 = \widehat{\tau^p}^2_{\bar{d}^1_k} + \widehat{\tau^p}^2_{\bar{d}^2_k}$ are estimated form the pilot study. Finally, the marginal power function is

$$\eta^m(n) = \int_0^\infty \eta(n; \tau^2)\pi(\tau^2|V^p_n)d\tau^2.$$

The sample size is selected as $\min\{n \in \mathbb{N} : \eta^m(n) > 1 - \beta\}$ for a given threshold $1 - \beta$.

3.3 Simulation study

In this section, we analyze through simulated data the sensitivity in terms of power and type I error of the proposed methodology and the existing frequentist sample size formulae for SMARTs to the misspecification of response rates, over-estimation of the standardized effect size or minimal detectable difference, and a breach of Assumption 3.1.

3.3.1 Framework

In this simulation study, we consider both continuous and binary outcomes, assuming the appropriateness of the Normal approximation in the latter case. Let p_{a_1} be the probability of response to the initial treatment a_1. If the final outcome is binary, $p_{a_1 a_2}$ is the probability of response to the second treatment a_2 when the first treatment a_1 fails, while subjects who have a positive reaction to the first stage intervention are considered as responders also at the second stage. If Y is continuous, the final outcome is sampled from a Normal distribution with mean $\mathrm{E}[Y|A_1, R, A_2]$ and variance $\mathrm{Var}[Y|A_1, R, A_2]$. Let us consider the SMART design illustrated in Figure 3.1. Following the same structure of Scott et al. [2007], we set $\mathrm{Var}[Y|A_1 = a_1, R = r, A_2 = a_2] = \zeta^2_{a_1,r,a_2}$, and we express the conditional mean as

$$\begin{aligned}\mathrm{E}[Y|A_1, R, A_2] = \phi_1 + \phi_2 I_{\{A_1=a\}} + \phi_3(1-R) + \phi_4 I_{\{A_1=a\}}(1-R) \\ + \phi_5 I_{\{A_2=c \,\cup\, A_2=e\}}(1-R) + \phi_6 I_{\{A_1=a \,\cup\, A_2=c\}}(1-R).\end{aligned}$$

It follows that the sets of parameters that need to be specified in the continuous outcome case are $\{\phi_l, l = 1, \dots, 6\}$ and $\{\zeta_{a,1,a}, \zeta_{a,0,c}, \zeta_{a,0,d}, \zeta_{b,1,b}, \zeta_{b,0,e}, \zeta_{e,0,f}\}$.

Scenarios

Following the SMART design outlined in Figure 3.1, we compare the strategies 'administer a and, if there is no response, switch to c' and 'administer b and, if there is no response, switch to e'. Note that these two treatment strategies are but two of the four possible strategies that are embedded within the trial. Varying the parameters of the data generating mechanism and the type of outcome, we simulate four scenarios:

- Scenario 1: the final outcome is continuous. The response rate to the first stage treatments is $p_a = p_b = 0.5$ and the sets of parameters used to generate the final outcome are

$$\{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\} = \{10, 5, -15, -3, 10, -3\},$$
$$\{\zeta_{a,1,a}, \zeta_{a,0,c}, \zeta_{b,1,b}, \zeta_{b,0,e}\} = \{2, 2, 2, 3\}.$$

  The resulting standardized effect size is 0.41 and the average difference between strategy means is 2.

- Scenario 2: continuous outcome. We set $p_a = p_b = 0.7$ and the simulation parameters of the outcome are

$$\{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\} = \{22, 5, -15, -7, 8, -3\},$$
$$\{\zeta_{a,1,a}, \zeta_{a,0,c}, \zeta_{b,1,b}, \zeta_{b,0,e}\} = \{6, 6, 2, 3\},$$

  resulting in a standardized effect size of 0.27 and a difference between strategy means of 2. Note that, in this scenario, Assumption 3.1 upon which the standard frequentist sample size formula relies does not hold.

- Scenario 3: binary outcome. The probabilities of response to the various treatments are $\{p_a, p_{ac}, p_b, p_{be}\} = \{0.3, 0.4, 0.3, 0.2\}$. The resulting standardized effect size and difference between strategy means are 0.28 and 0.14 respectively.

- Scenario 4: binary outcome. The probabilities of response are $\{p_a, p_{ac}, p_b, p_{be}\} = \{0.5, 0.65, 0.5, 0.5\}$, leading to the standardized effect size $\delta = 0.18$ and a difference between strategy means of 0.075. As in the second scenario, Assumption 3.1 does not hold.

Note that to calculate the effect size/minimal detectable difference from the simulation parameters in the continuous outcome case, we need to estimate $\mathrm{E}[Y|A_1, A_2]$ and $\mathrm{Var}[Y|A_1, A_2]$. The derivation of these quantities is carried out in Scott et al. [2007]. In the binary outcome case, this task is much simpler, as $\mathrm{E}[Y|A_1 = a_1, A_2 = a_2] = p_{a_1} + (1 - p_{a_1})p_{a_1a_2}$.

Model misspecification

To assess the robustness of the sample size estimates to the issue of local optimality, we consider two sources of misspecification: overestimation of the effect size or MDD and variability in the trial's parameters. Specifically, in the latter case, response rates to the first stage intervention are sampled from a Normal distribution truncated between 0 and 1 with mean p_{a_1} and standard deviation values from 0 to σ_m. If the outcome is binary, this variability is also added to the probabilities of success of the second treatment by sampling them from a truncated Normal distribution with mean $p_{a_1a_2}$. Moreover, we considered overestimations of the standardized effect size or minimal detectable difference by values up to 25%. In summary, for every scenario outlined in Section 3.3.1, the properties of the proposed Bayesian methodology are assessed for different choices of θ_0, σ_0^2 and σ_d^2 and under four combinations of the aforementioned sources of model misspecification:

- Setting 1: no misspecification.
- Setting 2: response rates present a standard deviation equal to σ_m.
- Setting 3: the minimal detectable difference is overestimated by 25%.
- Setting 4: combination of Settings 2 and 3.

The detailed data generating mechanism is outlined in Web Appendix B.

Additional specifications

Throughout this simulation study, we set the type I error α to 0.05 and the type II error β to 0.1. The standard deviation σ_m was set to 0.05 in Scenario 1, 0.04 in Scenarios 2 and 3, and 0.02 in Scenario 4. Using the calculations developed in Kim [2016], we determined the sample size of the simulated pilot studies (not to be confused with the full-scale trials) in order to guarantee that at least 6 individuals are observed in each treatment sequence with a 90% probability, which resulted in a sample size of 66 in Scenarios 1 and 4, 114 in Scenario 2, and 64 in Scenario 3. Finally, the type I error was assessed by sizing the full-scale trial to identify a minimal detectable difference between strategy means of 2 points in the continuous outcome case and 0.14 points in the binary outcome case when there is indeed no difference. The results are based on 3000 data replications.

3.3.2 Results

Partial results of the simulation study are presented below. Specifically, Table 3.1 shows the simulated power of the existing frequentist sample size formula in the continuous outcome case (Scenarios 1 and 2), whereas Tables 3.2-3.3 display the results relative to Scenario 1 under the proposed methodology. The full results of the simulation study and more details on the rationale behind the choice of these Scenarios are presented in Web Appendix C. Web Table S1 depicts the performance of the frequentist formula in terms of power in the binary outcome case and Web Table S2 summarizes the sample size estimates under this approach. Web Tables S3-S4, S5-S6, and S7-S8 show the results under the proposed Bayesian methodology in Scenarios 2, 3, and 4 respectively. Finally, Figure 3.2 presents a comparison in terms of power between the frequentist and the Bayesian formula, whereas Web Tables S9 and S10 show the simulated type I error in the continuous and binary outcome case respectively. The simulated power and type I error are estimated as the proportion of trials that identified a significant effect in the frequentist or Bayesian sense.

Power

As we can see from Table 3.1, the frequentist sample size formula performed well under the best-case scenario where there is no model misspecification and Assumption 3.1 is not violated (Scenario 1, top-left corner), nearing the desired 0.9 power level. However, its performance quickly deteriorates when the degree of model misspecification increases, causing power to fall to 0.82 when the standard deviation of the response rates reaches 0.05, 0.77 when the δ is overestimated by 25%, and 0.72 when both sources of model misspecification are present. In the second scenario we notice a similar trend, however, the decrease in power is more evident because of the violation of Assumption 3.1, which causes power to fall to 0.83 even in the absence of misspecification.

Table 3.1: Simulated power under the frequentist calculations in the continuous outcome scenarios, i.e. Scenarios 1 (left) and 2 (right), for different degrees of model misspecification.

		Scenario 1						Scenario 2				
		Response SD										
		0	0.01	0.02	0.03	0.04	0.05	0	0.01	0.02	0.03	0.04
% bias of δ	0	0.89	0.89	0.88	0.88	0.84	0.82	0.83	0.83	0.81	0.78	0.76
	5	0.88	0.87	0.86	0.84	0.82	0.81	0.81	0.80	0.78	0.77	0.74
	10	0.84	0.84	0.83	0.82	0.82	0.79	0.78	0.78	0.76	0.75	0.71
	15	0.82	0.83	0.81	0.80	0.79	0.76	0.74	0.75	0.73	0.72	0.70
	20	0.78	0.78	0.79	0.77	0.76	0.74	0.71	0.71	0.71	0.69	0.69
	25	0.77	0.76	0.76	0.75	0.72	0.72	0.68	0.69	0.66	0.65	0.64

On the other hand, the proposed Bayesian methodology provides us with the tools to offset the decrease in power caused by variability around response rates or overestimation of the standardized effect size/minimal detectable difference. Tables 3.2 and 3.3 provide the results of the simulation study under Scenario 1 when the mean of the analysis prior θ_0 is set to 0 and $\widehat{\theta}^p$ respectively. It is easily noticeable that the power level is independent of the choice of the analysis prior parameters θ_0 and σ_0, which, as expected, only affect the sample size.

Specifically, as σ_0 decreases, n increases under the neutral analysis prior centered at 0, and

Table 3.2: Power and average sample size (first and third quartile in brackets) under Scenario 1 (continuous outcome) computed using the Bayesian ‘two priors’ approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to 0.

		σ_d							
		0		0.2		0.5		0.8	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n
1	100	0.88	299 (263, 333)	0.89	312 (276, 347)	0.94	382 (335, 429)	0.99	593 (519, 663)
	3	0.89	311 (271, 346)	0.90	322 (283, 360)	0.94	395 (347, 439)	0.99	605 (528, 675)
	2	0.90	321 (283, 357)	0.90	333 (291, 373)	0.94	407 (357, 455)	0.99	617 (544, 687)
	1	0.88	366 (321, 410)	0.89	379 (332, 424)	0.94	462 (407, 514)	0.99	682 (599, 760)
2	100	0.83	301 (262, 336)	0.84	310 (270, 347)	0.88	381 (333, 425)	0.94	590 (516, 660)
	3	0.83	307 (269, 344)	0.82	319 (278, 355)	0.87	390 (342, 436)	0.94	598 (524, 667)
	2	0.82	318 (276, 357)	0.83	329 (287, 367)	0.87	401 (348, 450)	0.94	611 (536, 688)
	1	0.81	366 (322, 406)	0.81	379 (329, 425)	0.88	457 (397, 513)	0.94	676 (594, 753)
3	100	0.73	193 (169, 214)	0.76	198 (174, 220)	0.80	225 (196, 250)	0.87	289 (252, 322)
	3	0.74	201 (176, 224)	0.77	206 (181, 231)	0.81	233 (203, 260)	0.89	302 (265, 335)
	2	0.73	212 (185, 236)	0.75	217 (191, 243)	0.80	244 (214, 273)	0.87	313 (273, 350)
	1	0.72	254 (222, 283)	0.74	262 (228, 291)	0.81	293 (258, 326)	0.88	366 (319, 411)
4	100	0.72	192 (167, 213)	0.70	196 (171, 220)	0.74	224 (196, 249)	0.82	288 (252, 321)
	3	0.70	200 (175, 222)	0.70	206 (180, 229)	0.75	231 (201, 259)	0.81	298 (262, 332)
	2	0.69	211 (184, 235)	0.70	214 (187, 241)	0.76	245 (213, 274)	0.83	313 (273, 350)
	1	0.68	253 (223, 283)	0.67	258 (224, 288)	0.73	291 (254, 325)	0.81	366 (318, 411)

Table 3.3: Power and average sample size (first and third quartile in brackets) under Scenario 1 (continuous outcome) computed using the Bayesian 'two priors' approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to $\widehat{\mu}^p_{\bar{d}^1_k} - \widehat{\mu}^p_{\bar{d}^2_k}$ (estimated from the simulated pilot study).

		σ_d							
		0		0.2		0.5		0.8	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n
1	100	0.88	301 (264, 336)	0.89	311 (273, 345)	0.94	383 (337, 425)	0.99	592 (517, 662)
	5	0.89	293 (257, 328)	0.90	304 (264, 340)	0.94	374 (329, 417)	0.99	578 (507, 646)
	4	0.90	288 (251, 324)	0.90	298 (258, 334)	0.94	367 (321, 413)	0.99	564 (495, 632)
	3	0.88	275 (240, 311)	0.89	288 (251, 327)	0.94	352 (306, 397)	0.99	548 (478, 620)
2	100	0.82	299 (262, 335)	0.83	311 (272, 349)	0.88	381 (332, 429)	0.95	587 (516, 657)
	5	0.83	291 (253, 326)	0.83	300 (261, 337)	0.88	370 (325, 413)	0.94	567 (498, 636)
	4	0.82	286 (250, 322)	0.83	295 (257, 332)	0.88	363 (317, 407)	0.94	559 (487, 629)
	3	0.83	274 (237, 310)	0.84	283 (245, 322)	0.88	347 (301, 394)	0.93	541 (468, 611)
3	100	0.75	192 (168, 215)	0.75	197 (174, 219)	0.79	225 (196, 250)	0.88	291 (255, 325)
	5	0.76	187 (163, 209)	0.76	190 (167, 212)	0.81	217 (190, 243)	0.88	281 (247, 314)
	4	0.77	183 (159, 206)	0.78	187 (162, 210)	0.79	212 (184, 239)	0.88	277 (241, 309)
	3	0.76	175 (150, 199)	0.77	179 (154, 203)	0.80	204 (176, 230)	0.88	263 (226, 299)
4	100	0.71	191 (168, 213)	0.71	197 (173, 220)	0.73	223 (195, 249)	0.82	289 (253, 323)
	5	0.72	186 (164, 208)	0.72	189 (166, 211)	0.75	216 (189, 242)	0.82	278 (243, 312)
	4	0.73	183 (160, 205)	0.71	186 (161, 209)	0.76	212 (185, 238)	0.82	272 (236, 307)
	3	0.71	174 (149, 198)	0.72	177 (151, 202)	0.76	202 (173, 230)	0.82	261 (223, 297)

decreases if π_0 is centered at the difference between strategy means estimated via pilot data. When no variability around the minimal detectable difference is considered, i.e. $\sigma_d = 0$, the Bayesian formula generally leads to the same level of power of its frequentist counterpart in the four settings considered (which correspond to the 'corners' of Table 3.1), and similar average sample size values. However, the simulation results show how the addition of variability around the minimal detectable difference via the design prior π_d effectively mitigates the loss of power which affects the frequentist formula when the model is misspecified, generating sample size estimates that are more robust. Similarly, Web Tables S3 and S4 provide the results of the simulation study under the second scenario. As in Scenario 1, the increase of σ_d generates estimates that are more robust to the overestimation of the minimal detectable difference or variability around response rates. Additionally, since the Bayesian formula does not depend on Assumption 3.1, even if no variability around the minimal detectable difference is considered ($\sigma_d = 0$), it leads to a higher level of power with respect to the frequentist methodology, nearing the 0.9 level under no misspecification. Furthermore, in this specific setting, under a non-informative analysis prior the estimated average sample size is 741, which is 18% higher than the frequentist estimate, suggesting that the frequentist formula can potentially lead to underpowered studies when Assumption 3.1 is violated. Analogous considerations can be made in the binary final outcome case, as the performances of both the frequentist and Bayesian methods under Scenarios 3 and 4 respectively mirror the ones under Scenarios 1 and 2. Web Table S1 displays the performance of the frequentist formula in Scenarios 3 and 4, whereas the results related to the Bayesian methodology are presented in Web Tables S5-S6 and S7-S8. A partial representation of the power comparison between the two methods is depicted in the heatmaps of Figure 3.2, where the Bayesian methodology is assessed for a non-informative analysis prior and σ_d is set to 0.03.

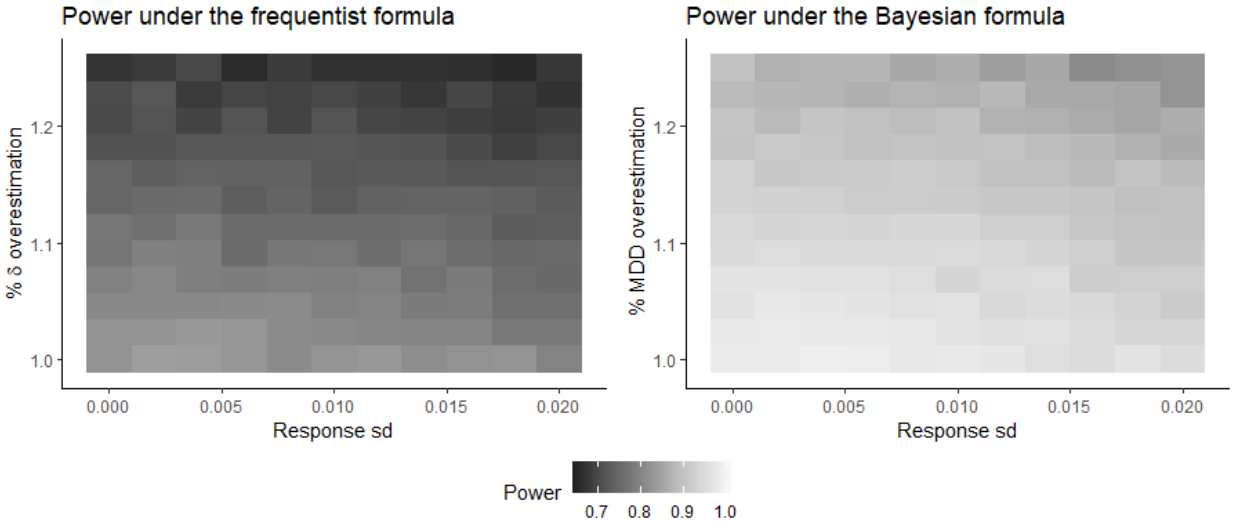

Figure 3.2: Heatmaps representing the simulated power under the frequentist and Bayesian methodology in Scenario 4 (binary outcome) for several degrees of model misspecification. For the Bayesian formula, the prior mean θ_0 is set to 0, $\sigma_0 = 100$ and $\sigma_d = 0.03$.

Finally, it is important to notice that the variability of the sample size estimates generated under the proposed methodology is higher in the scenarios where Assumption 3.1 does not hold. In fact, for example, considering the empirical distribution of the sample size estimates under each combination of prior parameters, under Scenario 4 the third quartile is on average 43% higher than the first quartile, whereas in Scenario 1 this difference amounts to 21%.

Type I error

Web Table S9 shows the sensitivity of type I error of the Bayesian formula under several prior specifications in the continuous outcome case. As expected, the simulated type I error generally attains the desired 0.05 level when θ_0 is set to $\widehat{\theta}^p$, and it decreases below that threshold as σ_0 decreases when the analysis prior π_0 is centered at 0. Web Table S10 provides the same information in the binary outcome case, and the results lead to the same conclusions.

3.4 Application to the Internet-Based Adaptive Stress Management SMART

The Internet-Based Adaptive Stress Management Pilot is a pilot SMART whose goal is to inform the planning of the subsequent larger full-scale study [Lambert et al., 2021]. The objective of this clinical trial is the evaluation of adaptive internet-based stress management interventions among adults with a cardiovascular disease and mild to severe levels of stress as measured by the Depression, Anxiety, and Stress Scale (DASS) [Lovibond and Lovibond, 1996]. The DASS is a set of self-reported scales aimed at the assessment of the level of depression, anxiety, and stress. Fourteen items are dedicated to each of the three conditions, resulting in three separate scores that range from 0 to 42. The stress scale evaluates difficulty to relax, nervous arousal, irritability, impatience, and agitation. Subjects with a score of 16 or higher were deemed eligible for this trial and, after 6 weeks of the first stage intervention, participants whose score fell below this threshold or improved by at least 50% with respect to their baseline assessment were considered as responders. Fifty-nine patients were enrolled and randomized to either a self-directed web-based stress management program or the same intervention with the addition of the assistance of a lay coach. In accordance with the SMART scheme outlined in Figure 3.1, after 6 weeks responders to the first stage intervention continued with the same program, whereas non-responders were randomized to their second stage interventions, which for both arms consisted of the continuation of the first treatment or the switch to a motivational interviewing based program. For the illustrative purposes of this section, we consider the two adaptive treatment strategies $\bar{d}_k^1 = \{$assign the 'website only' intervention and, if the patient does not respond, switch to the motivational interviewing based program$\}$ and $\bar{d}_k^2 = \{$assign the 'website + coach' intervention and, if the patient does not respond, switch to the motivational interviewing based program$\}$.

We size the full-scale version of the trial to test the hypothesis $\mu_{\bar{d}_k^1} - \mu_{\bar{d}_k^2} < 0$. More specifically, we compute the sample size to allow for the detection of a difference of 2 points in the DASS stress scale in favor of $\bar{d}_k^1$ with 90% power. Figure 3.3 shows the resulting power curves under different specifications of the prior parameters of π_0 and π_d and assuming the set of hyperparameters $(\theta_p, \kappa_p, \sigma_p^2, \nu_p) = (0, 1, 0.1, 5)$. Using a non-informative prior π_0 and setting $\sigma_d = 0$, the estimated sample size power is 490. If uncertainty around the minimal detectable difference is added through the standard deviation σ_d, the required sample size increases to 509 and 625 for $\sigma_d = 0.2$ and $\sigma_d = 0.5$ respectively. On the other hand, if we are willing to borrow further information from pilot data without accounting for uncertainty around the MDD (that is $\sigma_d = 0$), centering π_0 at the difference between strategy means estimated in the pilot study reduces the sample size to 455 if $\sigma_0 = 2$, 428 if $\sigma_0 = 1.5$, and 345 if $\sigma_0 = 1$.

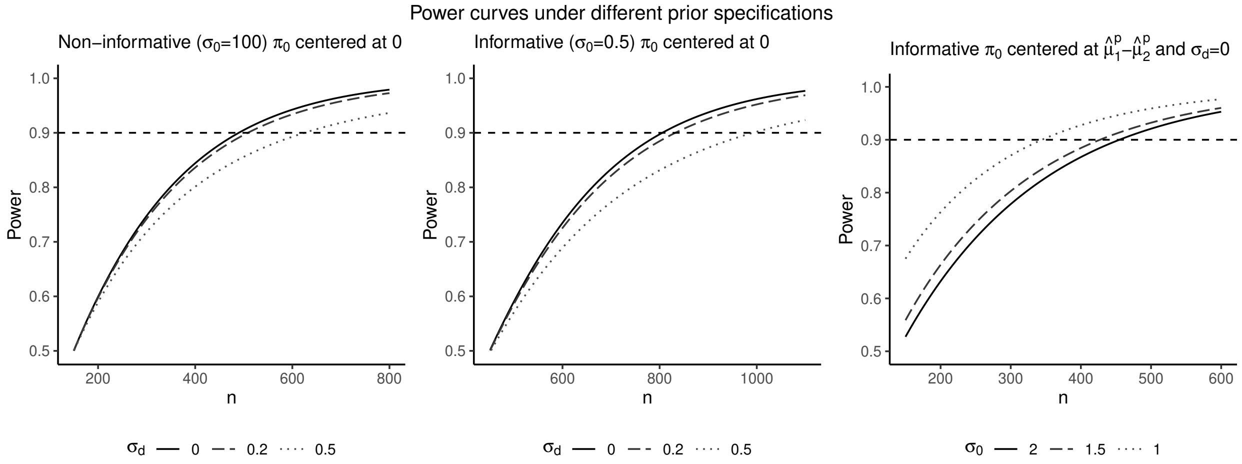

Figure 3.3: Power curves of the full-scale version of the Internet-Based Adaptive Stress Management SMART under different choices of prior parameters: non-informative analysis prior and varying standard deviation of the design prior (left), neutral informative analysis prior and varying standard deviation of the design prior (center), informative analysis prior centered at $\widehat{\theta}^p$ under different standard deviation values (right).

Since the frequentist approach requires additional assumptions on the initial treatments' response rates and is based on the specification of the standardized effect size rather than the MDD, a natural counterpart to the Bayesian sample size estimations is not achievable in real data applications. If Assumption 3.1 is not violated and setting a 40% probability of response to the first stage interventions, the sample size estimates under the frequentist approach for a standardized effect size of 0.20, 0.30, and 0.50 are 1372, 610, and 220 respectively.

3.5 Discussion

In this paper, we outlined a Bayesian extension to frequentist sample size formulae for SMARTs which relies on fewer assumptions and ensures more flexibility in the specification of key design parameters. The application of the ‘two priors’ approach to the framework of SMARTs allows us to (1) account for variability around the minimal detectable difference via the design prior π_d generating more reliable estimates and (2) integrate pre-trial knowledge via the analysis prior π_0. Furthermore, the marginalization of the Bayesian power function over the posterior distribution of the variance components estimated from pilot data ensures that the proposed methodology does not depend on the frequentist assumptions regarding the conditional variance of the outcome and the specification of intermediate response rates to initial treatments, which can be easily misspecified. Although pilot SMARTs are generally not sized to ensure precise estimates of the variance components, since the variability around them is encapsulated in their posterior distribution, this methodology is not compromised by the risk of underpowered full-scale trials that arises from the crude estimation of variance components from pilot data. Through a simulation study we demonstrated that, with respect to its frequentist counterpart, this methodology generally leads to sample size estimates that are more robust to model misspecification in terms of power. However, this procedure makes the sample size estimates subject to variability, and the simulation study showed that in certain scenarios the level of variability across data replications can be considerable. Moreover, the sample size estimates were on average higher than the frequentist estimations under a non-informative (or neutral) analysis prior. However, this increment is generally due to the greater assurance of the full-scale trial reaching the desired level of power that this methodology offers and, as we showed in the simulation study, it can be a consequence of the violation of the frequentist assumption on the conditional variance of the outcome, which can lead to underpowered full-scale trials under the frequentist estimates. Additionally, this method relies on the availability of pilot data for the estimation of the posterior distribution of τ^2. Although full-scale SMARTs are often preceded by a pilot study, postulating realistic values for τ^2 or the parameters of its distribution used to marginalize the Bayesian power function is a difficult task when such data are not available. In this case, our recommendation is the adoption of a simulation-based algorithm to estimate a plausible distribution for τ^2. One possible way to proceed is the following: using the same framework described in Section 3.3.1, hypothesize a set of realistic simulation parameters for the SMART study in question, iteratively generate data for its pilot study and collect the estimates of τ^2. Evaluate the empirical distribution of τ^2 obtained through this process and choose a set of parameters of the Noncentral chi-squared posterior distribution of τ^2 so that the resulting density is roughly consistent with the empirical density. It is advised to perform a sensitivity analysis employing different parameters for the simulation of the pilot study. Such parameters could also be treated as random variables and be sampled from an adequate distribution at each iteration to add more uncertainty to the process. Data from other studies with different designs might also be helpful to inform their plausible values. All the necessary R functions to carry out this procedure and implement the methodology outlined in this paper are available in the R package `bayesSMARTsize` accessible on Github.

Some limitations in connection with the choice of prior parameters need to be highlighted. First and foremost, it is crucial to recognize the different roles of the two priors. The analysis prior is the usual distribution used in Bayesian inference for the analysis stage. Borrowing pre-trial knowledge from pilot data through its elicitation can be a useful tool to decrease the sample size without compromising the power of the full-scale study, but its variance should be high enough to not overly influence the analysis stage of the full-scale trial. On other hand, the design prior formalizes uncertainty around the minimal detectable difference, hence, unlike the analysis prior, it must be a proper density with most of its mass lying in the alternative hypothesis space. The proposed methodology entails a certain level of subjectivity in the choice of hyperparameters. Although the shift of focus from the standardized effect size to the MDD might give a more straightforward course of action to elicit the prior distributions, we showed through the simulation study and the sizing of the Internet-Based Adaptive Stress Management SMART that sample size estimates and their properties vary substantially across different choices of hyperparameters. Therefore, a significant level of consideration and, eventually, a sensitivity analysis aimed at the selection of prior parameters are advised. Note that since the Bayesian and frequentist methodologies rely on different assumptions, it is not possible to achieve the frequentist estimates as a limiting case of the Bayesian approach. However, if Assumption 3.1 holds and the response rates to the initial treatments are correctly specified, for $\sigma_d^2 = 0$ and $\sigma_0^2 \to \infty$ the Bayesian methodology generates, on average, sample size values that are close to the frequentist estimates. Finally, in this paper, we chose a one-sided alternative hypothesis for illustrative purposes, but preliminary simulations under a two-sided alternative yielded comparable results (results not shown). The R package we have created allows the choice of the hypotheses type. Moreover, we focused on the simple SMART design with a continuous outcome where responders to the initial treatment are not re-randomized. A generalization of this methodology to other designs would require adjustments to the estimator of the strategy mean and its variance, but the Bayesian framework would remain generally similar. Furthermore, although we showed how the Normal approximation leads to satisfactory results when the final outcome is binary, the ad hoc extension of this approach to binary outcomes is an interesting avenue for further developments.

Acknowledgements

AT is supported by funds from a Recherche en Équipe grant from the FRQ – Nature et Technologie awarded to DAS. DAS and EEMM acknowledge support from Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). EEMM is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine and acknowledges the support of a chercheur de mérite career award from the FRQ – Santé.

Data availability statement

The data that support the findings in this paper are not shared as restrictions apply to the availability of these data, which were used under license in this paper.

Supporting Information

The Web Appendices and Tables referenced in Sections 3.2 and 3.3 are available with this paper at the Biometrics website on Wiley Online Library. The R package that implements the proposed methodology is available at https://github.com/aturchetta/bayesSMARTsize.

Chapter 4

A time-dependent Poisson-Gamma model for recruitment forecasting in multicenter studies

Preamble to Manuscript 2. This chapter presents the second manuscript of the thesis, which deals with the monitoring phase of multicenter clinical studies. One of the most popular methods developed to forecast recruitments in a multicenter clinical trial is the Poisson-Gamma model introduced by Anisimov and Fedorov [2007a, 2007b]. However, this model assumes that the underlying recruitment intensity remains constant over time, which is a scenario that is often not met in real clinical studies. The methodological contribution of this chapter consists of the development of a novel, flexible, time-dependent extension of this approach. The performance of the proposed methodology is assessed and compared to the standard Poisson-Gamma model in a simulation study and in a case study of a large multicenter cohort study.

Initially, the concept for this manuscript was to address the forecasting of recruitments in SMART studies. This idea stemmed from the (limited) evidence suggesting that, due to the appeal of SMARTs to study participants, the overall enrollment process might be more favorable than in standard RCTs. However, while the recruitment behaviors that were observed in SMART studies were included in the simulation study of this manuscript to validate the model, the developed approach is applicable to most types of multicenter study designs.

Note that this manuscript focuses on recruitment modeling, whereas the first manuscript presented in Chapter 3 focused on sample size determination. To ensure consistency with the well-established notation employed in these distinct research areas, the notation in this chapter is unrelated to that used in Chapter 3. For example, in the first manuscript α and β represented type I and II errors, whereas in this chapter they represent the shape and rate parameters of a Gamma distribution.

This manuscript was published in *Statistics in Medicine* [Turchetta et al., 2023].

A time-dependent Poisson-Gamma model for recruitment forecasting in multicenter studies

Armando Turchetta[1], Nicolas Savy[2], David A. Stephens[3], Erica E.M. Moodie[1], Marina B. Klein[4].

[1] *Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2] *Toulouse Mathematics Institute, University of Toulouse III*

[3] *Department of Mathematics and Statistics, McGill University*

[4] *Department of Medicine, Division of Infectious Diseases/Chronic Viral Illness Service, McGill University Health Center*

This thesis contains the accepted version of the corresponding paper published in *Statistics in Medicine* [Turchetta et al., 2023].

Abstract

Forecasting recruitments is a key component of the monitoring phase of multicenter studies. One of the most popular techniques in this field is the Poisson-Gamma recruitment model, a Bayesian technique built on a doubly stochastic Poisson process. This approach is based on the modeling of enrollments as a Poisson process where the recruitment rates are assumed to be constant over time and to follow a common Gamma prior distribution. However, the constant-rate assumption is a restrictive limitation that is rarely appropriate for applications in real studies. In this paper, we illustrate a flexible generalization of this methodology which allows the enrollment rates to vary over time by modeling them through B-splines. We show the suitability of this approach for a wide range of recruitment behaviors in a simulation study and by estimating the recruitment progression of the Canadian Co-infection Cohort.

4.1 Introduction

The recruitment time estimation is an important topic of interest in the planning and monitoring phase of multicenter studies with several practical implications. Yet, deterministic models mainly based on the study investigators' recruitment estimates are still used [Anisimov, 2008, Minois et al., 2017a]. This can lead to (1) ignoring important sources of variability and (2) overestimation of recruitment rates. The latter consequence is due to a phenomenon commonly known as the 'Lasagna Law' [Lasagna, 1979, van der Wouden et al., 2007], according to which medical investigators tend to be overly optimistic regarding the number of participants who meet the inclusion criteria and are willing to enroll in the study. These factors often lead to an underestimation of recruitment times and thus to difficulties in achieving the targeted sample size. For example, Walters et al. [2017] reviewed 151 RCTs conducted in the United Kingdom, concluding that 44% of the trials did not achieve the final recruitment target. A similar estimate was found by van der Wouden et al. [2007] analyzing 78 primary care research studies conducted in the Netherlands, who also noted that 51% of the studies had to extend the fieldwork period, and of these, 79% needed an extension longer than 50% of the originally planned study length.

A Bayesian approach built on a doubly stochastic Poisson process, known as the Poisson-Gamma model, was introduced to address the lack of a strong and consistent statistical methodology in this field in a series of papers by Anisimov & al. [Anisimov and Fedorov [2007a], Anisimov et al. [2007], Anisimov [2008], Anisimov [2009a], Anisimov [2011a]] This approach is based on modeling participants' arrival to recruitment centers as a Poisson process where the recruitment rates are assumed to be constant over time and to originate from a common Gamma distribution whose parameters are estimated from the ongoing trial in an empirical Bayes fashion. The model has been validated using several real trials' recruitment data and found to have good performances when the number of centers involved in the study is sufficiently high (>20). The Poisson-Gamma recruitment model has been further extended in numerous directions: Mijoule et al. [2012] considered the use of the Pareto distribution in place of the Gamma distribution, Bakhshi et al. [2013] and Minois et al. [2017a] suggested methods to exploit historical data from previous trials to estimate recruitment predictions before the start of the trial, Minois et al. [2017b] accounted for breaks in the recruitment process, and Anisimov et al. [2022] augmented the model to account for dropouts. The use of the Poisson process to capture the recruitment progression in clinical trials has been largely accepted in the literature and has deep roots. To the best of our knowledge, this framework was first introduced in Lee [1983], however, for a systematic review of early recruitment models for multicenter clinical trials we refer the reader to Gkioni et al. [2019]. More recently, Gajewski et al. [2008] modeled the waiting times between participants as exponential random variables (which is equivalent to a Poisson process) incorporating subjective knowledge on the recruitment process through an informative prior distribution, and Jiang et al. [2015] further elaborated this model by mitigating overly-optimistic investigators' assessments via adaptive priors.

All the aforementioned methods assume that the recruitment intensity remains constant over time. However, this assumption is seldom met in practice. A number of authors have proposed methods to mitigate this pitfall. Tang et al. [2012] proposed a discrete-time Poisson process-based method using a piece-wise linear function to model the accrual rate, where changepoints can be either fixed or estimated. However, this model was mostly tailored to their specific clinical trial example. A more general approach was introduced in Zhang and Long [2010], where the authors proposed a Bayesian method that relaxes the assumption of constant accrual rate via a non-homogeneous Poisson process where the overall underlying time-dependent accrual rate is modeled through cubic B-splines. This approach was further extended by Deng et al. [2017] in order to accommodate staggered initiation times and differences in accrual rates across regions. More recently, building on the standard Poisson-Gamma model, Lan et al. [2019] proposed a model where recruitment rates are assumed to be constant up to a certain point and then decay over time as a negative exponential. Urbas et al. [2022] further expanded this idea allowing for the detection of time-inhomogeneity via a testing procedure and considering a wider range of parametric curves to model the decaying recruitment evolution over time. All these methods performed well in the recruitment scenarios they were built to target.

In this paper, we present a novel extension of the Poisson-Gamma model which relaxes the constant-rate assumption. We model the recruitment process as a non-homogeneous Poisson process where the rates originate from a Gamma distribution whose mean parameter depends on time. Specifically, we assume that, at some unknown point in time, the recruitment progression will reach a plateau, and the non-constant section is modeled via B-splines; note that this approach is applicable to any multicenter study, whether a randomized trial or a cohort study. We outline the details of this methodology in Section 2. In Section 3, we assess in a simulation study the performance of the proposed approach in relation to the standard Poisson-Gamma model for a variety of recruitment behaviors. In Section 4, we apply our method to predict enrollments in the Canadian Co-infection Cohort. Section 5 concludes.

4.2 Methodology

Let $N_i(t)$ be the number of participants enrolled up to time t in center i, C the number of centers, $N(t) = \sum_{i=1}^{C} N_i(t)$ the total enrollments at time t, and u_i the center-specific initiation time.

4.2.1 The Poisson-Gamma model

The Poisson-Gamma (PG) model [Anisimov and Fedorov [2007a], Anisimov et al. [2007], Anisimov [2008], Anisimov [2009a], Anisimov [2011a]] assumes that participants arrive at each center according to a Poisson process with rate $\lambda_i(t) = I_{\{t>u_i\}}\lambda_i$. According to this model, the λ_i's are viewed as a sample from a Gamma distribution with parameters (α, β)

and probability density function

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}.$$

Hence, $N_i(t)$ is a Poisson process with cumulative rate

$$\Lambda_i(t) = I_{\{t>u_i\}}(t-u_i)\lambda_i$$

and, due to the superposition theorem of independent Poisson point processes, $N(t)$ is a Poisson process with cumulative rate

$$\Lambda(t) = \sum_{i=1}^{C} \Lambda_i(t) = \sum_{i=1}^{C} I_{\{t>u_i\}}(t-u_i)\lambda_i.$$

Consequently, assuming that we have reached a point in time t_{int} where all the C centers have started the enrollment phase, the posterior density of each rate λ_i is a Gamma distribution with parameters $\alpha + N_i(t_{int})$ and $\beta + (t_{int} - u_i)$.

The hyperparameters α and β can be estimated in an empirical Bayes fashion using the enrollment data collected up to t_{int}. Since the conditional distribution of $N_i(t_{int})$ is Poisson with rate parameter $(t_{int} - u_i)\lambda_i$, its marginal distribution is

$$N_i(t_{int}) \sim \text{NB}\left(\alpha, \frac{t_{int} - u_i}{\beta}\right)$$

where NB indicates a negative binomial random variable. Setting $m = \alpha/\beta$ and assuming that participants are recruited independently in the C centers, α and m can be estimated by maximizing the following likelihood:

$$\mathcal{L}(\alpha,\beta) = \prod_{i=1}^{C} \binom{\alpha + N_i(t_{int}) - 1}{\alpha - 1} \left(\frac{m(t_{int} - u_i)}{\alpha + m(t_{int} - u_i)}\right)^{N_i(t_{int})} \left(\frac{\alpha}{\alpha + m(t_{int} - u_i)}\right)^{\alpha}.$$

It follows that cumulative rate Λ is distributed as a sum of Gamma random variables, i.e.,

$$\widetilde{\Lambda} = \sum_{i=1}^{C} \text{Gamma}\left(\widehat{\alpha} + N_i(t_{int}),\, \widehat{\beta} + t_{int} - u_i\right).$$

Hence, the (additional) total number of enrollments $N^a(T)$ at time T is $N^a(T) = \text{Poisson}\left[\widetilde{\Lambda}(T - t_{int})\right]$. However, the constant-rate assumption is a restrictive limitation that is rarely appropriate for real applications in clinical studies. In the next section, we generalize the Poisson-Gamma model to the time dependence of the recruitment rates.

4.2.2 The time-dependent Poisson-Gamma model

Overview

We assume that participants arrive at each center according to a non-homogeneous Poisson process with rate $\lambda_i(t)$, where the rates are viewed as originating from a Gamma distribution whose mean parameter depends on time. Discretizing time in a pre-specified time unit, say days, let $n_i(t)$ be the number of enrollments on day t in center i, $N_i(t) = \sum_{s=1}^{t} n_i(s)$ the number of participants recruited up to day t in center i, $N(t) = \sum_{i=1}^{C} N_i(t)$ the total enrollments at time t across all centers, t_{int} the interim time when the predictions on future enrollments are estimated, and T the final point in time for which we estimate predictions. We assume that the center's recruitment intensity will plateau at an unknown point in time t_p. The main idea behind this methodology is to still employ the Poisson-Gamma recruitment model, but only using the recruitment data collected after the rates plateau. Each center initiates its recruitment process at time u_i. The initiation of the first center is set to day 1 ($u_1 = 1$), which is considered the start of the overall enrollment process. It follows that by time t, each initiated center has been recruiting participants for $t - u_i + 1$ days. In other words, the vector which represents the active recruiting days for center i is $\{1, 2, \ldots, t - u_i + 1\}$.

It is important to distinguish between the timeline of the overall recruitment process – which coincides with the timeline of the first initiated center – and the center-specific timeline. For example, assume that we have reached the interim time at day 200 of the recruitment process and the second center has been initiated on day 51, that is, $t = t_{int} = 200$, u_1=1, and $u_2 = 51$. We indicate with t^i_{int} the number of days each center has been recruiting until the interim time, i.e., $t^i_{int} = t_{int} - u_i + 1$. In this example, $t^1_{int} = 200$, $t^2_{int} = 150$, and the vector representing the recruiting days for the two centers are $\{1, 2, \ldots, 200\}$ and $\{1, 2, \ldots, 150\}$. Furthermore, suppose the plateau point is achieved at day 100 of the center's recruitment phase ($t_p = 100$). This means that we have observed 100 days of constant recruitment intensity for the first center and 50 days for the second center. Figure 4.1 summarizes the notation of the timelines.

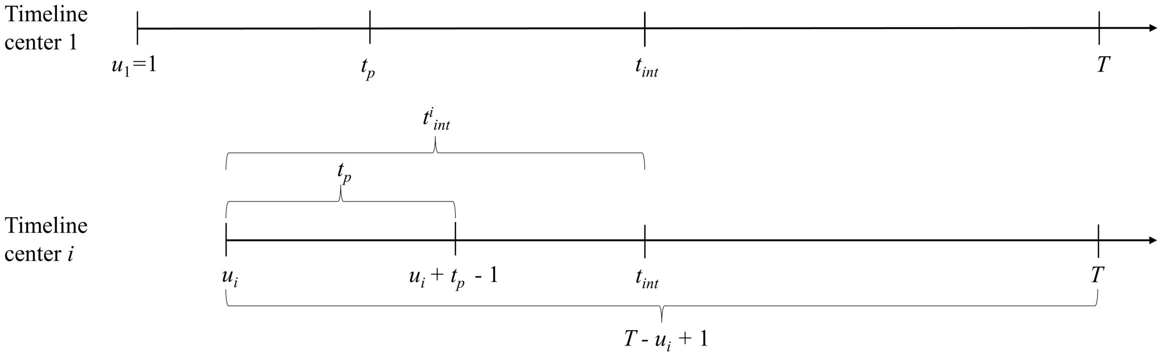

Figure 4.1: Timelines for the first (top) and i-th (bottom) center. Center 1: 1 is the start of recruitment for the entire study and corresponds to u_1, t_p represents the day the plateau begins, t_{int} the interim time, and T the final point in time for which predictions are estimated. Center i: u_i is its initiation time, the plateau for each center begins on day t_p of their recruitment process, which corresponds to day $u_i + t_p - 1$ of the overall recruitment phase, t^i_{int} is the number of days that center i enrolls participants from initiation to the interim time, and $T - u_i + 1$ is the total number of days during which center i can recruit, from initiation to the final time of prediction, T.

The time-dependent Poisson-Gamma model can be expressed as follows:

$$
\begin{aligned}
&N(t) \sim \text{Poisson}[\Lambda(t)],\\
&\Lambda(t) = \sum_{i=1}^{C} \Lambda_i(t) = \sum_{i=1}^{C} \int_{1}^{t-u_i+1} \lambda_i(s)ds = \sum_{i=1}^{C} \sum_{s=1}^{t-u_i+1} \lambda_i(s),\\
&\lambda_i(s) \begin{cases} \sim \text{Gamma}\left(\alpha, \frac{\alpha}{m(\boldsymbol{\phi}, s)}\right) & s \le t_p \\ = \lambda_i(t_p) & s > t_p \end{cases}
\end{aligned}
$$

where $\boldsymbol{\phi}$ is a set of parameters that needs to be estimated. Note that for $t_p = 1$ this model corresponds to the standard Poisson-Gamma model. We model the non-constant section of the rates' evolution over time through B-splines [De Boor, 1978]. Specifically, $m(\boldsymbol{\phi}, t)$ is modelled as follows:

$$
\log(m(\boldsymbol{\phi}, t)) = \begin{cases} \sum_{k=1}^{d} \eta_k \gamma_k(t) & t < t_p \\ \eta_d & t \ge t_p, \end{cases}
$$

where $\gamma_1(t), \ldots, \gamma_d(t)$ represent the basis functions.

Estimation

We separate the C centers into two groups, the first group containing the centers that have already passed the plateau point and the second group containing the ones that have not. Sorting the centers by initiation time, we indicate with C^* the number of centers that belong to the first group, that is, $C^* = \text{card}\{i : t_{int}^{i} \ge t_p\}$. Assuming that at time t_{int} all the centers have started the enrollment process, that is, $u_i < t_{int}\ \forall i$, the conditional distribution of the number of recruited participants in each center $N_i(t_{int})$ is Poisson with rate parameter

$\Lambda_i(t_{int}) = \sum_{s=1}^{t_{int}^i} \lambda_i(s)$, hence the conditional joint distribution of daily enrollments is

$$\begin{aligned}
\Pr\left[n_i(1),...,n_i(t_{int}^i)|\lambda_i(1),...,\lambda_i(t_{int}^i)\right] &\propto e^{-\sum_{s=1}^{t_{int}^i}\lambda_i(s)}\prod_{s=1}^{t_{int}^i}\lambda_i(s)^{n_i(s)} \\
&= \left[e^{-(t_{int}^i-t_p+1)\lambda_i(t_p)}\lambda_i(t_p)^{\sum_{s=t_p}^{t_{int}^i} n_i(s)}\right]^{1_{\{i\leq C^*\}}} \\
&\qquad e^{-\sum_{s=1}^{(t_p-1)\wedge(t_{int}^i)}\lambda_i(s)}\prod_{s=1}^{(t_p-1)\wedge t_{int}^i}\lambda_i(s)^{n_i(s)},
\end{aligned}$$

where $1_{\{i\leq C^*\}}$ is the indicator function which denotes the centers whose recruitment process has already passed the plateau point and $(t_p-1)\wedge t_{int}^i = \min\{(t_p-1), t_{int}^i\}$. Setting $N_i^*(t) = \sum_{s=t_p}^{t} n_i(s)$, the marginal predictive distribution is

$$\begin{aligned}
\Pr\left[n_i(1),...,n_i(t_{int}^i)\right] &= \int_0^\infty ... \int_0^\infty \Pr\left[n_i(1),...,n_i(t_{int}^i)|\lambda_i(1),...,\lambda_i(t_{int}^i)\right]\pi(\lambda_i(1),...,\lambda_i(t_{int}^i))\prod_{s=1}^{t_{int}^i} d\lambda_i(s) \\
&\propto \left[\frac{\Gamma(\alpha+N_i^*(t_{int}^i))}{\Gamma(\alpha)}\left(\frac{m(\boldsymbol{\phi},t_p)}{\alpha+m(\boldsymbol{\phi},t_p)}\right)^{N_i^*(t_{int}^i)}\left(\frac{\alpha}{\alpha+m(\boldsymbol{\phi},t_p)}\right)^{(t_{int}^i-t_p+1)\alpha}\right]^{1_{\{i\leq C^*\}}} \\
&\qquad \prod_{s=1}^{(t_p-1)\wedge t_{int}^i}\frac{\Gamma(\alpha+n_i(t_j))}{\Gamma(\alpha)}\left(\frac{m(\boldsymbol{\phi},s)}{\alpha+m(\boldsymbol{\phi},s)}\right)^{n_i(s)}\left(\frac{\alpha}{\alpha+m(\boldsymbol{\phi},s)}\right)^{\alpha}.
\end{aligned}$$

Finally, assuming independence between the number of participants recruited in different centers, the likelihood function is

$$\begin{aligned}
\mathcal{L}(\alpha,\boldsymbol{\phi}) \propto \prod_{i=1}^{C}&\left[\frac{\Gamma(\alpha+N_i^*(t_{int}^i))}{\Gamma(\alpha)}\left(\frac{m(\boldsymbol{\phi},t_p)}{\alpha+m(\boldsymbol{\phi},t_p)}\right)^{N_i^*(t_{int}^i)}\left(\frac{\alpha}{\alpha+m(\boldsymbol{\phi},t_p)}\right)^{(t_{int}^i-t_p+1)\alpha}\right]^{1_{\{i\leq C^*\}}} \\
&\qquad \prod_{s=1}^{(t_p-1)\wedge t_{int}^i}\frac{\Gamma(\alpha+n_i(t_j))}{\Gamma(\alpha)}\left(\frac{m(\boldsymbol{\phi},s)}{\alpha+m(\boldsymbol{\phi},s)}\right)^{n_i(s)}\left(\frac{\alpha}{\alpha+m(\boldsymbol{\phi},s)}\right)^{\alpha}.
\end{aligned}$$

Using this likelihood, α and $\boldsymbol{\phi} = \{t_p, \eta_1, \ldots \eta_d\}$ can be estimated and plugged into the posterior distribution of $\lambda_i(t)$.

Spline model choice

To guarantee enough flexibility in the recruitment curves that this model is able to estimate, we proposed to model the time-dependent enrollment intensity through B-splines. However, this choice implies that decisions have to be made regarding the degree of the polynomials, the number of knots, and their placement. Since these factors affect the dimension of the parameter space, we propose to fit different models and select the one which is best-fitting using the Bayesian Information Criterion (BIC) [Schwarz, 1978].

Forecasting future enrollments

For the centers that have already passed the plateau point, we can project the posterior distribution of $\lambda_i(t_{int})$ to predict future enrollments, so that the distribution of the additional recruited participants for center $i \in \{1, ..., C^*\}$ at time $T > t_{int}$ is Poisson with rate $\Lambda_i^a(T) = (T - t_{int})\lambda_i(t_{int})$, where

$$\lambda_i(t_{int}) \sim \text{Gamma}\left(\widehat{\alpha} + \sum_{s=\widehat{t}_p}^{t_{int}^i} n_i(s), \;\; \beta(t, \widehat{\boldsymbol{\phi}}) + t_{int}^i - \widehat{t}_p + 1\right), \qquad i = 1, ..., C^*,$$

and $\beta(t, \boldsymbol{\phi}) = \alpha / m(\boldsymbol{\phi}, t)$. On the other hand, for the remaining centers $i \in \{C^*+1, ..., C\}$,

$$\Lambda_i^a(T) = \sum_{s=t_{int}^i+1}^{\widehat{t}_p \wedge (T-u_i+1)} \lambda_i(s) + (T - u_i - \widehat{t}_p + 1)_+ \lambda_i(\widehat{t}_p),$$

$$\lambda_i(t) \sim \text{Gamma}\left(\widehat{\alpha}, \;\; \beta(t, \widehat{\boldsymbol{\phi}})\right),$$

where $(T - u_i - \widehat{t}_p + 1)_+ = \max(0, T - u_i - \widehat{t}_p + 1)$. Therefore, the distribution of future additional enrollments at time T, $N^a(T)$, is Poisson with overall recruitment rate

$$\Lambda^a(T) = \sum_{i=1}^{C^*}(T - t_{int})\lambda_i(t_{int}) + \sum_{i=C^*+1}^{C} \sum_{s=t_{int}^i+1}^{\widehat{t}_p \wedge (T-u_i+1)} \lambda_i(s) + \sum_{i=C^*+1}^{C} (T - u_i - \widehat{t}_p + 1)_+ \lambda_i(\widehat{t}_p).$$

It follows that the expectation and variance of the distribution of future enrollments are

$$\begin{aligned}\mathrm{E}[N^a(T)] &= \mathrm{E}\{\mathrm{E}[N^a(T)|\Lambda^a(T)]\} = \mathrm{E}[\Lambda^a(T)],\\ \mathrm{Var}[N^a(T)] &= \mathrm{Var}\{\mathrm{E}[N^a(T)|\Lambda^a(T)]\} + \mathrm{E}\{\mathrm{Var}[N^a(T)|\Lambda^a(T)]\} = \mathrm{Var}[\Lambda^a(T)] + \mathrm{E}[\Lambda^a(T)].\end{aligned}$$

The full expressions of the expected value and variance of $\Lambda^a(T)$ are displayed in the Appendix. Using these quantities, we can construct credible intervals (CrIs) for future enrollments using the Normal approximation. If the quantity of interest is the remaining time to reach the target number of recruitments, one can compute the expectations and credible intervals for future enrollments for a grid of values of T and then invert them to obtain a point estimate and credible interval of the remaining time to recruit the target number of participants.

4.3 Simulation study

In this section, we analyze the performance of the proposed time-dependent Poisson-Gamma model in a simulation study considering different types of curves for the underlying recruitment intensity and we compare it to the standard PG model. The model is assessed in terms of percentage bias and coverage rate by the 95% credible intervals of the observed future recruitments.

4.3.1 The generative model

The recruited number of participants at each time point t and for each center i is drawn from a Poisson distribution with rate $\widetilde{\lambda}_i(t)$, where $\widetilde{\lambda}_i(t)$ represents a draw from

$$\lambda_i(t) \sim \mathrm{Gamma}\left(\alpha, \frac{\alpha}{f^q(t)}\right), \quad i = 1, \ldots, C, \ t = 1, \ldots, t_p$$

and $\widetilde{\lambda}_i(t) = \widetilde{\lambda}_i(t_p)$ for $t > t_p$. The mean function $f^q(t)$ is the shifted scaled probability density function ($q = 1$) or the CDF ($q = 2$) of a Gamma random variable with various parameter choices, that is,

$$\begin{aligned} f^1(x) &= c_1 + c_2 \frac{p_2^{p_1}}{\Gamma(p_1)} e^{-p_2 x} x^{p_1 - 1}, \\ f^2(x) &= c_1 + c_2 \frac{p_2^{p_1}}{\Gamma(p_1)} \int_0^x e^{-p_2 u} u^{p_1 - 1} du. \end{aligned}$$

Varying the parameters of this generative model, we define five recruitment scenarios. The full list of parameters is provided in Table 4.1 and the resulting average recruitment curves are depicted in Figure 4.2. The parameters were selected in order to capture different common recruitment behaviors. The first two scenarios describe a slow start, and the use of the Gamma CDF to capture this evolution over time has been adopted in Deng et al. [2017] and Zhang and Long [2010]. In contrast, Scenarios 3 and 4 describe a fast start. We have included these enrollment scenarios after private discussions with project coordinators of studies where the recruitment process was aided by modern instruments such as social media. These studies experienced a fast start or an early bump in recruitment that was followed by the attainment of a plateau. Two of the studies where these behaviors were observed, for example, are the CanDirect study [McCusker et al., 2021] and the Canadian Co-infection Cohort study analyzed in Section 4. Finally, Scenario 5 describes the constant accrual setting that constitutes the framework of several recruitment models including the standard PG methodology. We evaluate the model for 20 and 60 centers in Setting 1 and

Table 4.1: Simulation parameters.

Scenario	c_1	c_2	p_1	p_2	t_p	q	α
1	0.2	0.50	0.55	0.09	60	1	1
2	0.2	0.50	10	0.15	150	1	1
3	0.2	8	1	0.05	130	2	1
4	0.2	13	2.40	0.11	100	2	1
5	0.2	0	NA	NA	1	NA	1

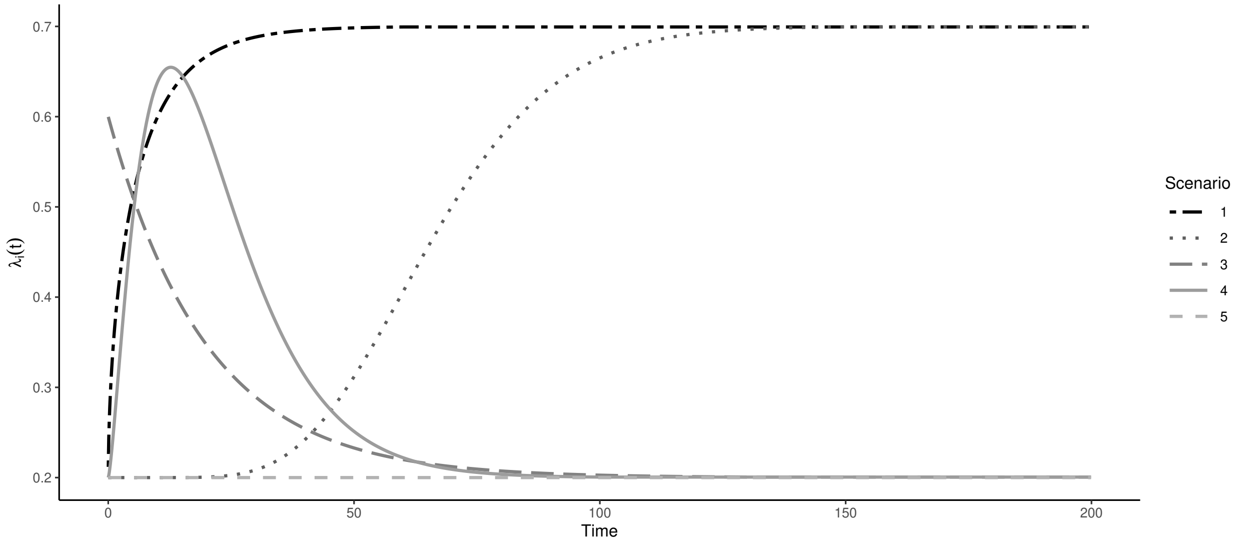

Figure 4.2: Average recruitment curves over time.

30 and 60 centers in Setting 2. We select two values of t_{int}, i.e., t_{int_1} and t_{int_2}, whose values depend on the scenario to assess the performance of the model when more information is collected.

Additionally, for each scenario, we consider two settings:

- Setting 1: all the centers start at the same time, i.e., $u_i = 1$ for all i.
- Setting 2: initiation times are staggered.

In Setting 2, we assume that the center-specific initiation times are distributed between $t = 1$ and the first interim time t_{int_1}, so that there are some centers whose recruitment intensity over time has not reached the plateau point. Two cases are considered: in Case 1, half of the centers have passed the plateau point ($C^* = \frac{1}{2}C$) by the first interim time t_{int_1}, whereas in Case 2 this proportion increases to two-thirds ($C^* = \frac{2}{3}C$). To ensure the consistency of these proportions across data replications, the initiation times of the first C^* centers are sampled from a Uniform distribution on $[1, t_{int_1} - t_p]$, and the remaining $C - C^*$ starting times from a Uniform density on $(t_{int_1} - t_p, t_{int_1})$. Therefore, by the time point t_{int_2}, some centers switch from the second group to the first one, i.e., C^* increases as $t_{int_1} \rightarrow t_{int_2}$. Given the structure of Setting 2, Scenario 5 is not considered as its plateau point is set at the start of the recruitment phase. In summary, we generated five scenarios in Setting 1 and four in Setting 2. For each combination of scenarios and settings, we consider two interim times, two values of C, and two cases for Setting 2.

As for the choice of spline models for $m(\boldsymbol{\phi}, t)$, we fitted four B-splines models varying the degree of the polynomials and the number of knots. To limit the dimension of the vector of parameters, we considered quadratic and cubic splines and one and zero internal knots, where the placement of the knot was set to $\widehat{t}_p/2$. The best-fitting model among the four candidates was selected using the BIC. The results are based on 1000 data replications.

4.3.2 Results

Table 4.2 illustrates the performance of the proposed time-dependent Poisson-Gamma (tPG) model and the standard PG technique in Setting 1 in terms of coverage rate of the observed recruitments by the 95% credible interval (CrI), percentage bias, and standard error (SE). Table 4.3 provides the same information for the setting with staggered initiation times.

Table 4.2: Coverage rate (CR) of the 95% CrI, percentage bias, and standard error (SE) over 1000 data replications under the proposed method (tPG) and the standard Poisson-Gamma (PG) model in Setting 1 (same initiation time across centers).

	Scenario	T	t_{int}	tPG		PG	
				C		C	
				20	60	20	60
CR	1	300	80	0.93	0.95	0.27	0.24
			160	0.94	0.96	0.48	0.47
	2	500	200	0.94	0.95	0.07	0.01
			300	0.95	0.95	0.12	0.01
	3	500	160	0.90	0.92	0.16	0.04
			250	0.94	0.95	0.28	0.08
	4	400	120	0.88	0.90	0.02	0
			200	0.96	0.94	0.06	0
	5	300	80	0.96	0.94	0.95	0.95
			160	0.95	0.96	0.94	0.93
Bias	1	300	80	4.98	2.93	14.05	8.53
			160	2.98	1.57	7.30	4.40
	2	500	200	3.31	1.89	22.65	22.78
			300	2.31	1.33	15.16	15.28
	3	500	160	7.92	4.53	30.72	26.41
			250	4.45	2.53	19.89	16.89
	4	400	120	9.96	5.43	62.64	55.76
			200	4.77	2.96	37.72	33.52
	5	300	80	5.25	3.08	5.33	2.99
			160	4.71	2.72	4.88	2.76
SE	1	300	80	660	1187	211	375
			160	431	758	280	493
	2	500	200	907	1591	247	441
			300	610	1064	315	558
	3	500	160	276	507	92	156
			250	220	388	115	201
	4	400	120	230	417	86	146
			200	179	306	97	164
	5	300	80	204	356	203	353
			160	128	224	126	222

In general, the tPG model led to a significant improvement with respect to the PG method in Scenarios 1-4 in terms of CrIs coverage rates and percentage bias. As expected, the standard Poisson-Gamma model failed to deliver acceptable results, as it generated biased estimates in the four scenarios where the recruitment intensity is time-dependent, with the percentage bias nearing or exceeding 50% and the CrIs coverage rate reaching 0 in several instances. On the other hand, the time-dependent Poisson-Gamma model was able to capture the non-constant evolution of the recruitment intensity over time providing more accurate predictions.

Additionally, in Scenario 5, the tPG and PG methods led to similar performances, indicating the suitability of the proposed methodology even in the scenario where the assumptions of the PG model are satisfied.

Overall, Table 4.2 shows that the tPG model performed well in every scenario being considered in Setting 1, as it led to adequate levels of percentage bias and coverage rates which improved as the number of centers increments or more information on the recruitment process is accrued by moving forward the interim time t_{int}. However, when the initiation times are staggered, the proposed model's performance decreased. Specifically, the model leads to lower levels of coverage rates of the observed recruitments by the credible intervals, and the percentage bias can reach values over 10%. This decline in forecasting accuracy is more evident in Scenarios 3 and 4 when the proportion of centers whose recruitment intensity has not reached the plateau is the highest, i.e. in Case 1 for the first interim time. The model's performance greatly improves as the proportion of centers that belong to C^* increases either by switching from Case 1 to Case 2 or by delaying the interim time.

Table 4.3: Coverage rate (CR) of the 95% CrI, percentage bias, and standard error (SE) over 1000 data replications under the proposed method (tPG) and the standard Poisson-Gamma (PG) model in Setting 2 (staggered initiation times).

				tPG				PG			
	Scenario	T	t_{int}	Case 1		Case 2		Case 1		Case 2	
				C=30	C=60	C=30	C=60	C=30	C=60	C=30	C=60
CR	1	300	80	0.88	0.91	0.91	0.92	0.05	0.02	0.09	0.03
			110	0.93	0.94	0.94	0.93	0.05	0.01	0.07	0.03
	2	500	200	0.87	0.85	0.89	0.86	0	0	0	0
			260	0.90	0.90	0.92	0.92	0	0	0	0
	3	500	160	0.73	0.71	0.84	0.83	0.20	0.17	0.18	0.14
			220	0.89	0.89	0.92	0.92	0.31	0.32	0.30	0.31
	4	400	120	0.77	0.74	0.80	0.75	0.03	0	0.06	0.01
			170	0.87	0.86	0.89	0.86	0.10	0.03	0.13	0.06
Bias	1	300	80	12.12	8.43	10.53	7.15	26.84	27.79	22.44	22.85
			110	8.35	5.75	6.92	4.82	25.04	25.67	20.58	20.88
	2	500	200	13.44	9.51	10.57	7.74	53.88	54.86	46.41	47.49
			260	10.29	7.53	8.28	5.79	43.97	44.82	37.99	38.82
	3	500	160	15.97	11.93	12.05	8.98	22.22	16.75	23.80	19.44
			220	9.39	7.02	7.58	7.60	13.59	9.54	13.96	9.98
	4	400	120	16.14	12.37	14.54	10.91	52.13	48.12	47.23	41.80
			170	9.80	7.48	8.50	6.27	32.00	28.79	29.78	25.78
SE	1	300	80	708	967	731	1003	270	358	217	298
			110	679	949	677	958	220	300	242	340
	2	500	200	946	1498	1011	1484	198	276	196	270
			260	802	1212	855	1212	244	367	273	392
	3	500	160	311	453	299	429	210	282	149	220
			220	291	399	288	417	98	132	99	141
	4	400	120	312	514	308	462	159	216	193	258
			170	245	385	258	364	96	132	93	128

Finally, it is of interest to look at which models were selected by the BIC. By and large, the BIC showed an overwhelming preference for smaller models with no internal knots in Scenarios 1, 2, and 5, whereas in Scenarios 3 and 4 models that include an internal knot were chosen more frequently.

4.4 Forecasting recruitments in the Canadian Co-infection Cohort study

In this section, we utilize the time-dependent Poisson-Gamma model to estimate enrollments in the Canadian Co-infection Cohort (CCC) study. The CCC is a prospective observational study whereby participants living with both HIV and Hepatitis C (HCV) are recruited and monitored through follow-up visits scheduled every 6 months [Klein et al., 2010]. The primary goal of the study is to achieve a better understanding of the risk factors associated with liver disease and its progression in the growing population of HCV–HIV co-infected people, with particular emphasis on the effect of highly active antiretroviral therapy (HAART) and HCV treatment. The CCC study encompasses 19[1] recruitment centers across Canada opened between March 2003 and June 2014 and, as of April 8th, 2022, it comprises a total of 2077 recruited participants in 995 weeks.

To assess the applicability and accuracy of the proposed methodology in forecasting the enrollments in this study, we considered three interim times t_{int}. At each of these time points, the accrued data are used to estimate the recruitment progression up to April 8th, 2022, i.e. week 995, and the resulting forecasts are compared to the observed enrollments. Since 13 of the 19 centers were opened between week 192 (November 2006) and 285 (September 2008), we set the interim times to $t_{int} = 400, 500, \text{and } 600$ weeks. The results under the proposed methodology and the standard Poisson-Gamma model are illustrated in Figure 4.3. As we can see from the observed recruitment progression, the constant-rate assumption is not met. Specifically, the centers generally showed a higher recruitment intensity in the initial stages of their enrollment period, followed by a deceleration until the attainment of a plateau, which our model consistently estimated at 110 weeks after the centers' opening for every interim time. The violation of the constant-rate assumption is especially visible after the opening of the cluster of centers between weeks 192 and 285, as the recruitment progression

[1] Two centers were merged but are still treated separately in this analysis.

visibly decelerates around week 330. Regardless of the interim time, the tPG model led to an estimated recruitment progression that closely tracks the observed enrollments and, apart from brief time periods immediately following the interim times, the resulting 95% credible intervals covered the observed enrollments. Note that the last two years of recruitment were affected by the COVID-19 pandemic, whose start is identified in Figure 4.3 as March 11th, 2020, i.e. the day the WHO declared the COVID-19 outbreak a global pandemic. The pandemic caused a deceleration in enrollments, and its beginning coincides with the period where the point estimates produced by the tPG model start to deviate the most from the observed recruitments, especially for $t_{int} = 500$.

On the other hand, the forecasts under the standard Poisson-Gamma model are heavily biased. More specifically, since the centers showed an initial fast recruitment intensity followed by a slowdown, the standard PG model led to an overestimation of the recruitment rates and hence to overly-optimistic forecasts.

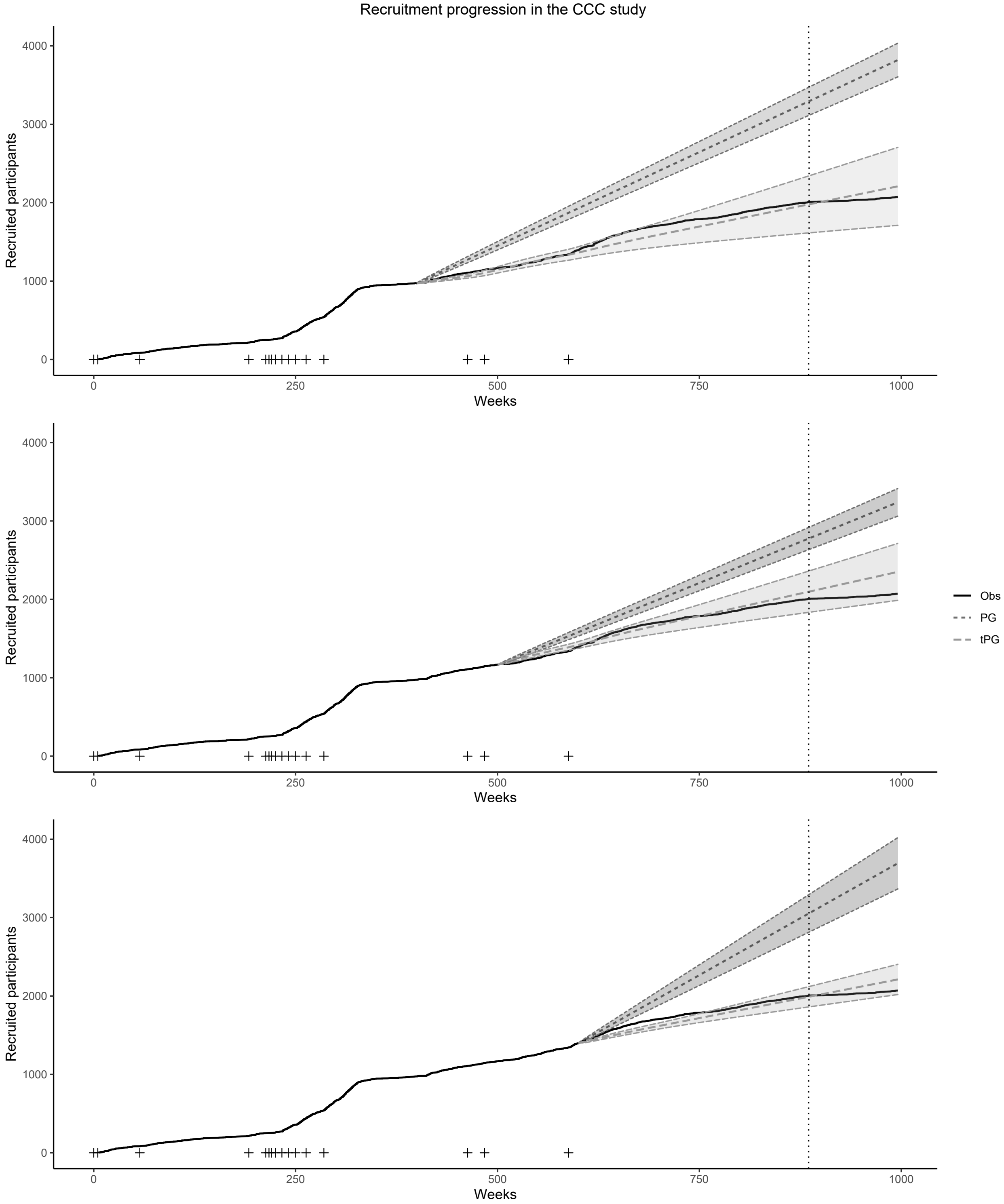

Figure 4.3: Observed and estimated recruitments under the time-dependent Poisson-Gamma (tPG) and Poisson-Gamma (PG) models in the CCC study at three interim times ($t_{int} = 400, 500, \text{and } 600$ weeks). The shaded regions represent the 95% credible intervals, the '+' signs the opening date for each center, and the dotted vertical line the start of the COVID-19 pandemic.

4.5 Discussion

In this paper, we have introduced an extension to the Poisson-Gamma recruitment model where the recruitment intensity is allowed to vary over time. Specifically, with respect to the standard methodology where recruitment rates are assumed to be constant, we introduced the presence of a time window that spans from the centers' initiation to an unknown time point where the recruitment intensity is time-dependent. In order to guarantee the applicability of this methodology for a wide range of recruitment scenarios, we modeled the mean parameter of the common Gamma distribution which generates the recruitment rates through B-splines. We analyzed the performance of the proposed methodology in a simulation study for various behaviors of the recruitment curves. The model performed well in every scenario being considered when the centers share the same initiation time. On the other hand, the setting where the centers' initiation times are staggered showed a decline in performance for some scenarios. Specifically, the proposed model struggled to deliver overall satisfactory results when a sizable number of centers has not reached the plateau point in their enrollment process, especially for the more elaborate recruitment behaviors assumed in Scenarios 3 and 4. The model's performance greatly improved when the proportion of these centers was reduced or the interim time was delayed (hence still decreasing this proportion), which suggests that while the model correctly estimates the constant section of the recruitment curves, it may not adequately catch the non-constant section in some scenarios. Note that due to the computational time and the variety of settings considered, we limited the number of spline models to four. Adding more models to the list of candidates by varying the number of internal knots or/and their location might improve the model's accuracy. This is partially confirmed by the BIC model choice. In fact, the BIC selected mostly smaller models to estimate the simpler recruitment curves, whereas larger models with an internal knot were preferred more frequently in the scenarios characterized by more intricate recruitment progressions. However, even in the scenarios where the model's performance could be improved, the proposed recruitment model still led to significantly better results than the standard Poisson-Gamma model. Additionally, in the scenario where the constant-intensity assumption is met, the time-dependent Poisson-Gamma model led to comparable results, suggesting that the proposed methodology can be seen as a time-dependent generalization of the standard PG model without any evident loss in performance.

The application of the tPG model to the recruitment data of the Canadian Co-infection Cohort study showcased its suitability and accuracy in forecasting enrollments in a real study even for a low number of recruitment centers (16 for the first interim time), while also highlighting the limitations of the standard PG model when the constant-rate assumption is not met.

Finally, some other limitations of the methodology we have presented in this paper need to be highlighted. In particular, this model relies on the availability of recruitment data past the plateau point for an adequate number of centers, which makes it impractical to employ in the planning phases of a clinical study or during its early monitoring stage. The integration of prior information from similar studies or the centers' history is an interesting topic for further research. Moreover, we assumed that the plateau point is the same for all centers. The addition of a layer of variability to reflect the variations across centers' could represent a relevant improvement in the flexibility of this methodology.

The `tPG` R package developed to implement the proposed recruitment model is available on GitHub.

Acknowledgements

AT is supported by a doctoral fellowship from the Fonds de Recherche du Québec (FRQ) - Nature et Technologie and by funds from a Recherche en Équipe grant from the FRQ – Nature et Technologie awarded to DAS. DAS and EEMM acknowledge support from Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). EEMM is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine and acknowledges the support of a chercheur de mérite career award from the FRQ - Santé. MBK is supported by a Tier I Canada Research Chair in Clinical and Epidemiologic Studies of Chronic Viral Infections in Vulnerable Populations. The Canadian Coinfection Cohort is funded by the Réseau sida/maladies infectieuses du Fonds de Recherche du Québec—Santé; the Canadian Institute for Health Research (CIHR; FDN-143270); and the CIHR Canadian HIV Trials Network (CTN222).

Data availability statement

The data that support the findings in this paper are not shared as restrictions apply to the availability of these data, which were used under license in this paper. The R package that implements the proposed methodology is available at https://github.com/aturchetta/tPG.

APPENDIX

Expected value and variance of $N^a(T)$

The expected value and variance of the additional number of future recruited participants $N^a(T)$ can be expressed as

$$\begin{aligned}
\mathrm{E}[N^a(T)] &= \mathrm{E}\{\mathrm{E}[N^a(T)|\Lambda^a(T)]\} = \mathrm{E}[\Lambda^a(T)], \\
\mathrm{Var}[N^a(T)] &= \mathrm{Var}\{\mathrm{E}[N^a(T)|\Lambda^a(T)]\} + \mathrm{E}\{\mathrm{Var}[N^a(T)|\Lambda^a(T)]\} = \mathrm{Var}[\Lambda^a(T)] + \mathrm{E}[\Lambda^a(T)],
\end{aligned}$$

where

$$
\begin{aligned}
\mathrm{E}[\Lambda^a(T)] &= \sum_{i=1}^{C^*}(T-t_{int})\mathrm{E}[\lambda_i(t_{int})] + \sum_{i=C^*+1}^{C}\sum_{s=t^i_{int}+1}^{\widehat{t}_p\wedge(T-u_i+1)}\mathrm{E}[\lambda_i(s)] + \sum_{i=C^*+1}^{C}(T-u_i-\widehat{t}_p+1)_+\mathrm{E}[\lambda_i(\widehat{t}_p)] \\
&= (T-t_{int})\sum_{i=1}^{C^*}\frac{\widehat{\alpha}+\sum_{s\in\{\widehat{t}_p,\ldots,t^i_{int}\}}n_i(s)}{\beta(t,\widehat{\boldsymbol{\phi}})+t^i_{int}-\widehat{t}_p} + \sum_{i=C^*+1}^{C}\sum_{s=t^i_{int}+1}^{\widehat{t}_p\wedge(T-u_i+1)}\frac{\widehat{\alpha}}{\beta(s,\widehat{\boldsymbol{\phi}})} \\
&\quad + \sum_{i=C^*+1}^{C}(T-u_i-\widehat{t}_p+1)_+\frac{\widehat{\alpha}}{\beta(\widehat{t}_p,\widehat{\boldsymbol{\phi}})},
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[\Lambda^a(T)] &= \sum_{i=1}^{C^*}(T-t_{int})^2\mathrm{Var}[\lambda_i(t_{int})] + \sum_{i=C^*+1}^{C}\sum_{s=t^i_{int}+1}^{\widehat{t}_p\wedge(T-u_i+1)}\mathrm{Var}[\lambda_i(s)] \\
&\quad + \sum_{i=C^*+1}^{C}(T-u_i-\widehat{t}_p+1)^2_+\mathrm{Var}[\lambda_i(\widehat{t}_p)] \\
&= (T-t_{int})^2\sum_{i=1}^{C^*}\frac{\widehat{\alpha}+\sum_{s\in\{\widehat{t}_p,\ldots,t^i_{int}\}}n_i(s)}{(\beta(t,\widehat{\boldsymbol{\phi}})+t^i_{int}-\widehat{t}_p)^2} + \sum_{i=C^*+1}^{C}\sum_{s=t^i_{int}+1}^{\widehat{t}_p\wedge(T-u_i+1)}\frac{\widehat{\alpha}}{\beta(s,\widehat{\boldsymbol{\phi}})^2} \\
&\quad + \sum_{i=C^*+1}^{C}(T-u_i-\widehat{t}_p+1)^2_+\frac{\widehat{\alpha}}{\beta(\widehat{t}_p,\widehat{\boldsymbol{\phi}})^2}.
\end{aligned}
$$

Chapter 5

The time-dependent Poisson-Gamma model in practice: recruitment forecasting in HIV trials and a tutorial

Preamble to Manuscript 3. The previous chapter introduced the time-dependent Poisson-Gamma model for forecasting enrollments in multicenter studies, and the approach was validated on the recruitment data from a cohort study. However, the use of statistical models to forecast recruitments is possibly of more practical relevance in randomized clinical trials where there is a target sample size to be achieved within a given time frame.

The manuscript presented in this chapter has two aims: (1) to assess the proposed model's accuracy on recruitment data from randomized clinical trials, and (2) to illustrate an easy to follow tutorial on its implementation via the `tPG` R package. The second aim stemmed from a recent paper by Gkioni et al. [2020]. The authors surveyed chief investigators and statisticians involved in the planning and monitoring phase of the recruitment stage of clinical studies, concluding that statistical models to forecast recruitments were rarely used. The main reasons cited for the avoidance of statistical models to predict recruitments were the non-familiarity with either the models or their implementation and the simplicity of a deterministic model. This chapter takes a more practical route compared to the other two manuscripts, hoping to facilitate the implementation of the proposed recruitment model in practice. The notation in this manuscript is consistent with that of the previous manuscript (Chapter 4).

The manuscript presented in this chapter will be submitted to a statistical journal soon after the submission of this thesis.

The time-dependent Poisson-Gamma model in practice: recruitment forecasting in HIV trials and a tutorial

Armando Turchetta[1], Erica E.M. Moodie[1], David A. Stephens[2], Nicolas Savy[3], Zoe Moodie[4], Janine van Duijn[5], Wouter Willems[6].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2]*Department of Mathematics and Statistics, McGill University*

[3]*Toulouse Mathematics Institute, University of Toulouse III*

[4]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center*

[5]*Janssen Vaccines Research & Prevention BV, Janssen Pharmaceutical Companies of Johnson & Johnson*

[6]*Janssen Research & Development, Janssen Pharmaceutical Companies of Johnson & Johnson*

Abstract

Despite a growing body of literature in the area of recruitment modeling for multicenter studies, in practice, statistical models to predict enrollments are rarely used and when they are, they often rely on unrealistic assumptions. The time-dependent Poisson-Gamma model (tPG) is a recently developed flexible methodology which allows analysts to predict recruitments in an ongoing multicenter trial, and its performance has been validated on data from a cohort study. In this article, we illustrate and further validate the tPG model on recruitment data from randomized controlled trials and provide a practical and easy to follow guide to its implementation via the `tPG` R package. To validate the model, we show the predictive performance of the proposed methodology in forecasting the recruitment process of two HIV vaccine trials conducted by the HIV Vaccine Trials Network in multiple Sub-Saharan countries.

5.1 Introduction

Recruitment forecasting for multicenter clinical studies has become a popular topic in recent years. The successful recruitment of the desired number of participants in a clinical trial is a long-standing issue [Bieganek et al., 2022], and several methods have been introduced to address this problem. Given the nature of the enrollment phase of a clinical study, the Poisson process is the most straightforward statistical tool to model the participants' arrival process. To the best of our knowledge, Lee [1983] was the first author to formalize a statistical framework for the forecasting of recruitments in clinical trials using the Poisson distribution. The underlying assumption needed to use a Poisson process to model recruitments in multicenter studies is that the sites enroll participants at the same rate, however, in practice, there can be substantial heterogeneity across centers. To address this shortcoming, Anisimov and coauthors [2007a, 2007, 2008, 2009a, 2011a] developed the doubly-stochastic Poisson-Gamma (PG) recruitment model, which adds to the Poisson process a second layer of variability by introducing a Gamma prior distribution for the center-specific recruitment rates. This approach has been validated on several data from clinical trials [Anisimov and Fedorov, 2005, Anisimov, 2009b, Zhang and Huang, 2022] and augmented to allow for the use of the Pareto distribution in place of the Gamma density [Mijoule et al., 2012], unknown initiation times [Anisimov, 2009a], breaks in the recruitment process [Minois et al., 2017b], and participants' drop-out either upon arrival or during the screening period [Anisimov et al., 2022]. Furthermore, Turchetta et al. [2023] introduced a time-dependent extension of the Poisson-Gamma model whereby the recruitment rates are allowed to vary with time from their initiation up to the attainment of a plateau point, and the model was validated on recruitment data from a large cohort study. In fact, the assumption that recruitment rates are constant over time can be restrictive and is often not met in practice, which motivated the development of a growing number of time-dependent recruitment models. Zhang and Long [2010] developed a non-homogenous Poisson process whereby the overall recruitment intensity is modeled via a cubic B-spline function up to the monitoring time when predictions are estimated, and is assumed to be constant thereafter. Deng et al. [2017] extended this approach to accommodate staggered centers' initiation times and region-specific differences in recruitment rates. Lan et al. [2019] introduced a simulation-based time-dependent extension to the Poisson-Gamma model where the recruitment rates are considered constant up to an assumed point in time and then decay as a negative exponential. This approach was augmented in Urbas et al. [2022] by including a test for the detection of time-inhomogeneity and different candidate parametric curves to capture the decay in recruitment intensity.

In practice, statistical approaches to forecast enrollments are rarely used, and deterministic approaches based on the study investigators' assumption on the centers' recruitment rates are significantly more common [Gkioni et al., 2020]. This raises a series of concerns, such as the omission of important sources of variability and the well-known propensity of investigators to elicit overly optimistic recruitment rates, which is a phenomenon known as Lasagna Law [Lasagna, 1979, Bogin, 2022]. This disconnect between the scientific literature and its implementation is particularly worrisome in light of the overwhelming evidence which states that (1) a significant proportion of studies fails to achieve their original sample size target within the planned time frame or at all [Walters et al., 2017, Jacques et al., 2022], and (2) poor recruitment is one of the leading causes of trial discontinuation [Kasenda et al., 2014]. While the consequences of the first point may range from the extension of the recruitment period (and hence a greater financial burden) to the conduct of an underpowered study, discontinuing a clinical trial raises ethical concerns over the depletion of scientific resources, since the results of such studies are rarely reported. When Gkioni et al. [2020] surveyed 69 statisticians involved in planning and monitoring the recruitment phase of clinical trials, 90% stated that they did not use any statistical model, citing the simplicity of a deterministic model, the non-familiarity with either the available statistical methods or their implementation, and doubts over the additional value of these methods as the main reasons.

This paper has two main goals:

1. to illustrate and validate the time-dependent Poisson-Gamma model recently introduced in Turchetta et al. [2023] on recruitment data from randomized controlled clinical trials, and
2. to provide a practical guide to its implementation via the newly developed `tPG` R package.

To validate the model, we obtained recruitment data from two HIV vaccine trials conducted in Sub-Saharan Africa by the HIV Vaccine Trials Network. Recruitment is particularly challenging in HIV vaccine studies, as participants face several barriers to entry such as stigma or discrimination due to participation in an HIV trial and fear of vaccine-induced HIV infection [Allen et al., 2001, Andrasik et al., 2020].

The paper is structured as follows: Section 2 presents an overview of the time-dependent Poisson-Gamma model. In Section 3, we validate the model on the recruitment data from the HIV vaccine trials and compare its predictive performance to the standard Poisson-Gamma model. In Section 4, we offer an easy to follow tutorial on the functions embedded in the `tPG` R package to estimate the model parameters and forecast the recruitment process. Section 5 concludes.

5.2 The time-dependent Poisson-Gamma model

Let C be the number of sites involved in a multicenter clinical trial, $N_i(t)$ the number of participants enrolled up to time t in center i, $N(t) = \sum_{i=1}^{C} N_i(t)$ the total enrollments at time t, and u_i the center-specific initiation time.

Model overview

According to the time-dependent Poisson-Gamma (tPG) model, the study participants arrive at each recruitment center independently and according to a non-homogeneous Poisson process with time-dependent rates $\lambda_i(t)$. The recruitment rates are assumed to originate from a Gamma distribution with rate parameter α and shape $\alpha/m(\boldsymbol{\phi}, t)$, so that the resulting mean parameter $m(\boldsymbol{\phi}, t)$ depends on time. The main assumption regarding the recruitment intensity is that at an unknown time point t_p, or plateau point, the enrollment rates will stabilize and be constant until the end of the recruitment period, but up until that point they are allowed to vary with time. The non-constant portion of $\log(m(\boldsymbol{\phi}, t))$ is modelled via B-splines in order to ensure sufficient flexibility in terms of the number of recruitment scenarios the model is able to capture. Essentially, this creates a 'buffer zone' between the center's initiation date and t_p where the recruitment intensity can have a wide range of behaviors. Indicating with $\Lambda_i(t)$ and $\Lambda(t)$ the cumulative center-specific and overall recruitment rates, if we discretize time in a pre-specified time unit, e.g., days, the tPG model can be summarized as follows:

$$\begin{aligned}
&N(t) \sim \text{Poisson}[\Lambda(t)],\\
&\Lambda(t) = \sum_{i=1}^{C} \Lambda_i(t) = \sum_{i=1}^{C} \sum_{s=1}^{t-u_i+1} \lambda_i(s),\\
&\lambda_i(s) \begin{cases} \sim \text{Gamma}\left(\alpha, \frac{\alpha}{m(\boldsymbol{\phi}, s)}\right) & s \leq t_p \\ = \lambda_i(t_p) & s > t_p \end{cases}
\end{aligned}$$

where

$$\log(m(\boldsymbol{\phi}, t)) = \begin{cases} \sum_{k=1}^{d} \eta_k \gamma_k(t) & t < t_p \\ \eta_d & t \geq t_p, \end{cases}$$

and $\boldsymbol{\phi}$ is a set of parameters that needs to be estimated which includes the B-spline parameters $\eta_1, \ldots, \eta_d$ and the plateau point t_p, while $\gamma_1(t), \ldots, \gamma_d(t)$ represent the basis functions, and d remains to be chosen. Like the standard Poisson-Gamma model, the tPG model is an empirical Bayes approach. As such, once an interim or monitoring time t_{int} is reached, the parameters α and $\boldsymbol{\phi}$ can be estimated from the ongoing trial after computing the likelihood function for the number of enrollments. For more details on the computation, see Turchetta et al. [2023]. Throughout the rest of this paper, we use days as the time unit, but weeks or months could be valid alternatives depending on the characteristics of the study in question.

As for the choice of B-spline model, we suggest fitting several candidate models varying the number and placement of internal knots between the center's initiation and its stabilization point. Since that is unknown, the placement can be defined as a function of the estimated plateau point. The best-fitting model is then selected via the Bayesian information criterion (BIC) [Schwarz, 1978]. In the case studies analyzed in the Section 5.3, eight candidate models were fitted at every interim time varying the polynomial degree (quadratic and cubic) and the placement of one internal knot (none, $\widehat{t_p}/2$, $\widehat{t_p}/3$, $\widehat{t_p}/4$).

Predictions

Once the parameters of the tPG model are estimated and the BIC-preferred model selected, the parameters are plugged in the distribution of the recruitment rates. The idea is to include only the enrollment data accrued during the constant recruitment phase into the distribution of the rates. Hence, the distribution of the individual rates varies depending on whether the centers have passed the plateau point by the interim time. Let us sort the centers by initiation date and indicate with C^* the number of centers that have already passed this point by the interim time. For these centers, indicating with t_{int}^i the number of days each center has been recruiting until the interim time and with $n_i(t)$ the number of enrollments for center i on day t, the posterior distribution of the recruitment rates at time t_{int} is

$$\lambda_i(t_{int}) \sim \text{Gamma}\left(\widehat{\alpha} + \sum_{s=\widehat{t}_p}^{t^i_{int}} n_i(s),\ \ \beta(t,\widehat{\boldsymbol{\phi}}) + t^i_{int} - \widehat{t}_p + 1\right), \qquad i = 1, ..., C^*,$$

where $\beta(t, \boldsymbol{\phi}) = \alpha/m(\boldsymbol{\phi}, t)$. This distribution is then projected until the end of the trial to predict the number of additional future enrollments at time T, which will have a Poisson distribution with cumulative rate $\Lambda^a_i(T) = (T - t_{int})\lambda_i(t_{int})$, where the superscript a over the cumulative rate or the number of recruited patients indicates that these quantities refer to the additional recruitments from the interim time until T rather than the total enrollments of the trial. For the remaining centers $i \in \{C^* + 1, ..., C\}$, the distribution of the recruitment rates at time t is $\lambda_i(t) \sim \text{Gamma}\left(\widehat{\alpha}, \beta(t, \widehat{\boldsymbol{\phi}})\right)$, and the number of future additional recruitments at time T is Poisson distributed with cumulative rate

$$\Lambda^a_i(T) = \sum_{s=t^i_{int}+1}^{\widehat{t}_p \wedge (T-u_i+1)} \lambda_i(s) + (T - u_i - \widehat{t}_p + 1)_+ \lambda_i(\widehat{t}_p),$$

where $(T - u_i - \widehat{t}_p + 1)_+ = \max(0, T - u_i - \widehat{t}_p + 1)$.

Finally, summing up these components, it follows that the distribution of total future additional enrollments at time T, i.e., $N^a(T)$, is Poisson with overall recruitment rate

$$\Lambda^a(T) = \sum_{i=1}^{C^*} (T - t_{int})\lambda_i(t_{int}) + \sum_{i=C^*+1}^{C} \sum_{s=t^i_{int}+1}^{\widehat{t}_p \wedge (T-u_i+1)} \lambda_i(s) + \sum_{i=C^*+1}^{C} (T - u_i - \widehat{t}_p + 1)_+ \lambda_i(\widehat{t}_p).$$

In order to make predictions using this model, one could either use Monte Carlo simulations or rely on the Normal approximation to compute point estimates and credible intervals (CrIs). Given the efficacy and simplicity of the second route, we prefer the second option. In fact, using the basic properties of conditional expectation and variance, it follows that

$$\begin{aligned}\mathrm{E}[N^a(T)] &= \mathrm{E}\{\mathrm{E}[N^a(T)|\Lambda^a(T)]\} = \mathrm{E}[\Lambda^a(T)],\\ \mathrm{Var}[N^a(T)] &= \mathrm{Var}\{\mathrm{E}[N^a(T)|\Lambda^a(T)]\} + \mathrm{E}\{\mathrm{Var}[N^a(T)|\Lambda^a(T)]\} = \mathrm{Var}[\Lambda^a(T)] + \mathrm{E}[\Lambda^a(T)].\end{aligned}$$

If the target quantity of interest is the remaining time to recruit a desired number of participants, the expectation and credible interval for the future enrollments are computed for a grid of values $T > t_{int}$ and then inverted. Figure 5.4 displays a visual representation of this process.

Inclusion of covariates

If there are systematic differences in the recruitment intensity across centers which can be explained by center-specific characteristics such as, for example, the size or region of the recruitment sites, it might be worth incorporating such differences into the recruitment model. A straightforward course of action would be to add covariates to the mean parameter of the Gamma distribution, which would then depend on the center i and take the form

$$\log(m_i(\boldsymbol{\phi}, t)) = \begin{cases} \sum_{k=1}^{d} \eta_k \gamma_k(t) + \sum_{j=1}^{r} x_{i,j}\zeta_j & t < t_p \\ \eta_d + \sum_{j=1}^{r} x_{i,j}\zeta_j & t \geq t_p, \end{cases}$$

where ζ_j represents the coefficient of covariate x_j. As more data are accrued, the weight of the data in the posterior distribution of the rates overshadows the prior parameters, hence this inclusion of covariates can have an impact on predictions only when a center has not passed the plateau point or the site has not yet been initiated (i.e., the center has not yet begun recruiting). However, in our simulations (results not shown), including covariates did not lead to a predictive improvement if the distribution of covariates is independent of the initiation times, despite simulating differences in recruitment intensity due to center-specific characteristics. The only scenario where such inclusion did lead to a visible improvement is when the correlation is present, the center-effect is large, and enough centers have not passed the plateau point or been initiated by the interim time. For example, this situation may occur if centers in urban areas recruit participants at a higher rate than rural sites and are also more likely to be initiated later in time. Figure 5.1 depicts an illustrative example of this scenario. Data were simulated assuming that urban centers enroll participants three times faster than rural sites. In the first plot, urban and rural centers' initiation times are random, whereas in the second one urban sites are more likely to be initiated later in time.

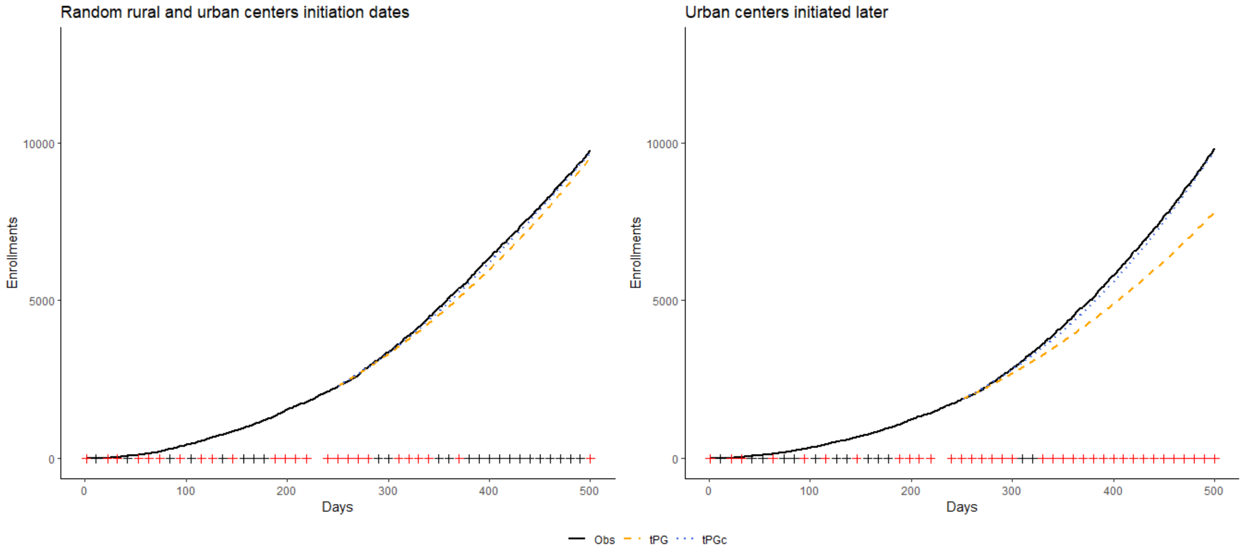

Figure 5.1: Recruitment forecastings in simulated data sets under the time-dependent PG model. The model with covariates is indicated with tPGc. The + marks represent the initiation dates for rural (black) and urban (red) centers.

5.3 Forecasting enrollments in two clinical trials from the HIV Vaccine Trials Network

In this section, we forecast the enrollments in two clinical trials conducted in Sub-Saharan Africa by the HIV Vaccine Trials Network (HVTN) which seek to evaluate the safety and efficacy of two HIV vaccine candidates. A counterexample where both the time-dependent and standard Poisson-Gamma models fail to capture the observed recruitment progression is presented in the appendix via the analysis of the Promotion of Breastfeeding Intervention Trial (PROBIT) data [Kramer et al., 2001].

5.3.1 HVTN 702

The HVTN 702 study (ClinicalTrials.gov identifier: NCT02968849), also known as Uhambo, was a randomized, double-blind, placebo-controlled, phase 2b–3 trial which sought to evaluate the safety and efficacy of an experimental preventive HIV vaccine regimen consisting of the ALVAC-HIV vaccine + Bivalent Subtype C gp120 protein adjuvanted with MF59 [Gray et al., 2021]. The trial was sponsored by the National Institute of Allergy and Infectious Diseases (NIAID) and it was conducted in South Africa across 14 recruitment centers. Between October 2016 and June 2019, 5407 healthy adults were randomized in a 1:1 ratio to the vaccine regimen or placebo. The participants were monitored over time, with the primary outcome being the occurrence of HIV-1 infection (the most common type of HIV) within 24 months from randomization. Unfortunately, although the planned sample size of 5400 participants had been met, vaccinations in the trial were halted in January 2020 when an interim analysis met the pre-specified criteria for non-efficacy. After inspecting the overall and center-specific recruitment progression, we noticed that there were three yearly breaks during the December holiday period. To adjust for this, we removed 10 days between the end of December and start of January each year from the recruitment data, hence the number of days in either the observed process or predictions has to be interpreted as 'effective recruitment days'. After this adjustment, the total number of days of the recruitment period amounts to 944.

To validate the time-dependent Poisson-Gamma model on the recruitment data from this clinical trial, we set four interim times at 250, 400, 550, and 700 days after the initiation of the first center. At each monitoring time, the data collected up to that point are used to estimate the model parameters, and the resulting predictions on future enrollments and the time needed to achieve the target sample size are compared to the observed recruitment process and the results obtained under the standard PG model. The results are shown in Table 5.1 and Figure 5.2.

Table 5.1: Estimated time (in effective days) to achieve the final sample size (5407) and predicted enrollments at the end of the observed recruitment period (day 944) in the HVTN 702 trial under the time-dependent and standard Poisson-Gamma models for different interim times. The number of activated centers C_{act} and percentage of participants already recruited by the time predictions are computed are also shown.

	t_{int}	% recruited	C_{act}	Time (95% CrI)	Recruitments (95% CrI)
tPG	250	18%	13	945 (725, 1781)	5405 (3133, 7677)
	400	40%	13	763 (649, 1238)	7094 (4378, 9809)
	550	59%	13	911 (803, 1317)	5612 (4478, 6746)
	700	79%	14	865 (825, 974)	5952 (5309, 6594)
PG	250	18%	13	1018 (969, 1074)	4968 (4686, 5249)
	400	40%	13	857 (837, 880)	6036 (5858, 6214)
	550	59%	13	863 (847, 880)	5989 (5850, 6128)
	700	79%	14	863 (852, 874)	5975 (5879, 6071)

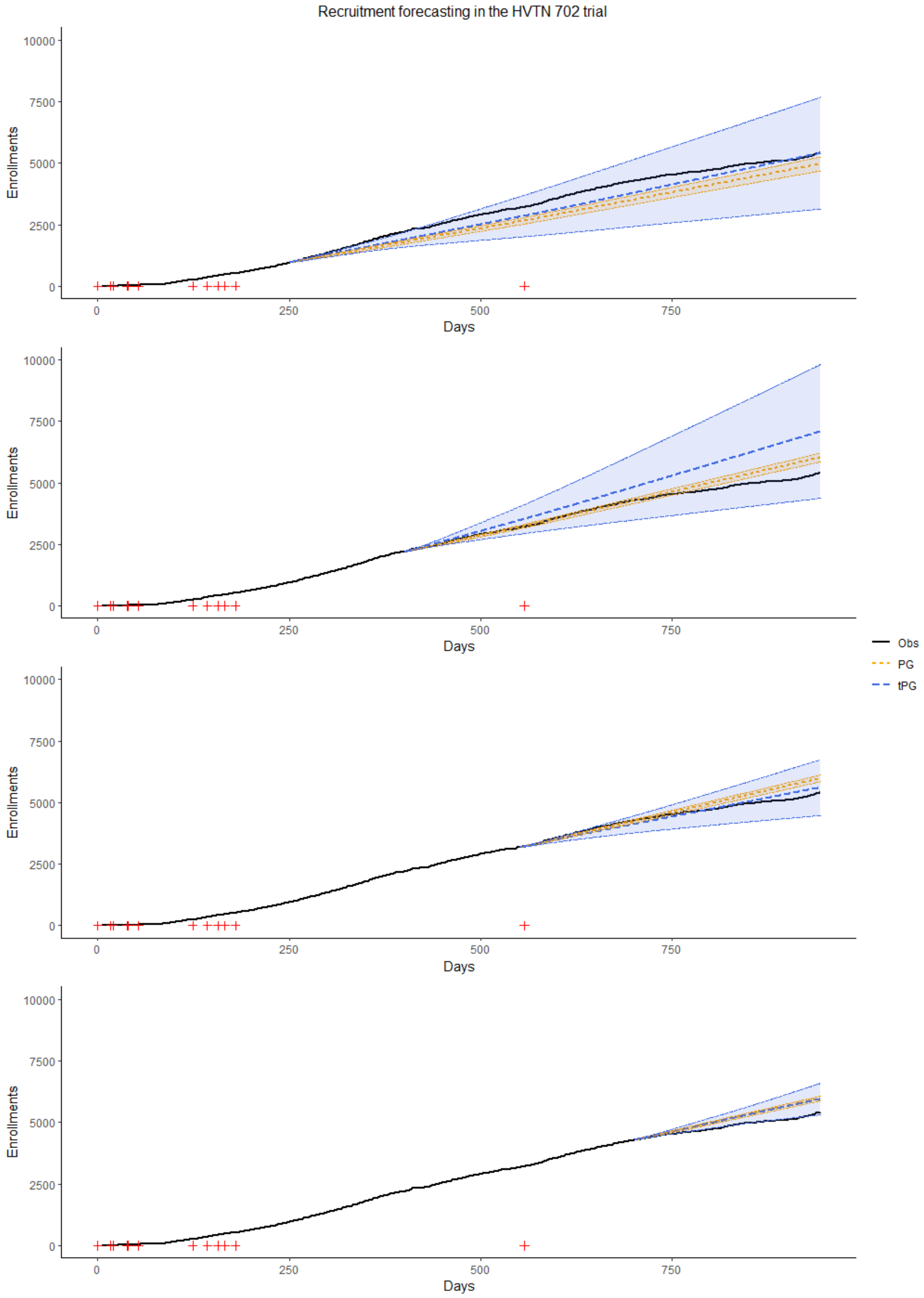

Figure 5.2: Enrollment predictions in the HVTN 702 study under the time-dependent and standard PG models compared to the observed recruitments at four interim times ($t_{int} = 250, 400, 550,$ and 700 days). The shaded areas indicate the respective 95% credible intervals and the + marks represent the centers' initiation dates.

As we can see, for any interim time considered, the most notable difference between the standard and time-dependent Poisson-Gamma model is in the width of the credible intervals (CrIs). The standard PG model delivers CrIs that are overly narrow and fail to include the observed recruitments. On the other hand, the time-dependent PG model leads to much larger and realistic credible intervals that do include the observed recruitment process, whether in terms of the total enrollments at the end of the trial or the estimated time to achieve the desired sample size. Interestingly, under the tPG model, while the width of the credible intervals of the remaining time to complete the enrollment phase decreases as the monitoring time moves forward, the CrI of the total enrollments at the end of the trial increases from the first to the second interim time. Albeit counterintuitive, this is possible when the estimated plateau point moves forward in time as more data are accrued. This was the case for this study, as the tPG model estimated stabilization points very close to the interim time, indicating that the recruitment intensity was indeed non-constant for the majority of the enrollment phase. Additionally, since several candidate models were fitted at each interim time, the model selected by the BIC varied at the different time points. More details on the estimated parameters are laid out in the appendix.

In terms of point estimates, the differences between the two models are less evident. Compared to the PG model, the tPG approach leads to point estimates that are closer to the observed recruitments for the first and second interim times, further away when $t_{int} = 400$, and almost identical for $t_{int} = 750$.

5.3.2 HVTN 705

The HVTN 705 trial (ClinicalTrials.gov identifier: NCT03060629), also known as Imbokodo, was a randomized, double-blind, placebo-controlled, phase 2b trial sponsored by Janssen Vaccines & Prevention B.V. and conducted in five Sub-Saharan African countries [ClinicalTrials.gov]. The goal of the trial was to evaluate the safety and efficacy of a novel vaccine regimen (Ad26.Mos4.HIV and aluminum-phosphate adjuvanted Clade C gp140) in preventing HIV-1 infections in women. In the course of 575 days from October 2017 to May 2019, 2637 HIV-seronegative women between the age of 18 and 35 were enrolled in 23 sites spread across South Africa, Mozambique, Malawi, Zimbabwe, and Zambia. The primary endpoint of the study was the number of HIV-1 infections diagnosed between 7 and 24 months after the administration of the first vaccination dose. Unfortunately, this study failed to show a significant protective effect of the candidate vaccine regimen [Kenny et al., 2022]. Similarly to HVTN 702, this study was characterized by a visible slowdown in recruitment around the December holiday period. However, in this case some centers continued enrolling during this period, hence we did not remove those days. Given the shorter duration of the recruitment process, we set the monitoring times at 150, 300, and 450 days after the initiation of the first center. The results are shown in Table 5.2 and Figure 5.3.

Table 5.2: Estimated time (in days) to achieve the final sample size (2637) and predicted enrollments at the end of the observed recruitment period (day 575) in the HVTN 705 trial under the time-dependent and standard Poisson-Gamma models for different interim times. The number of activated centers C_{act} and percentage of participants already recruited by the time predictions are computed are also shown.

	t_{int}	% recruited	C_{act}	Time (95% CrI)	Recruitments (95% CrI)
tPG	150	5%	7	603 (511, 824)	2415 (1489, 3340)
	300	22%	13	573 (536, 635)	2664 (2224, 3103)
	450	57%	20	595 (572, 632)	2475 (2277, 2673)
PG	150	5%	7	770 (698, 872)	1649 (1387, 1911)
	300	22%	13	659 (630, 695)	2090 (1934, 2247)
	450	57%	20	622 (607, 641)	2318 (2233, 2402)

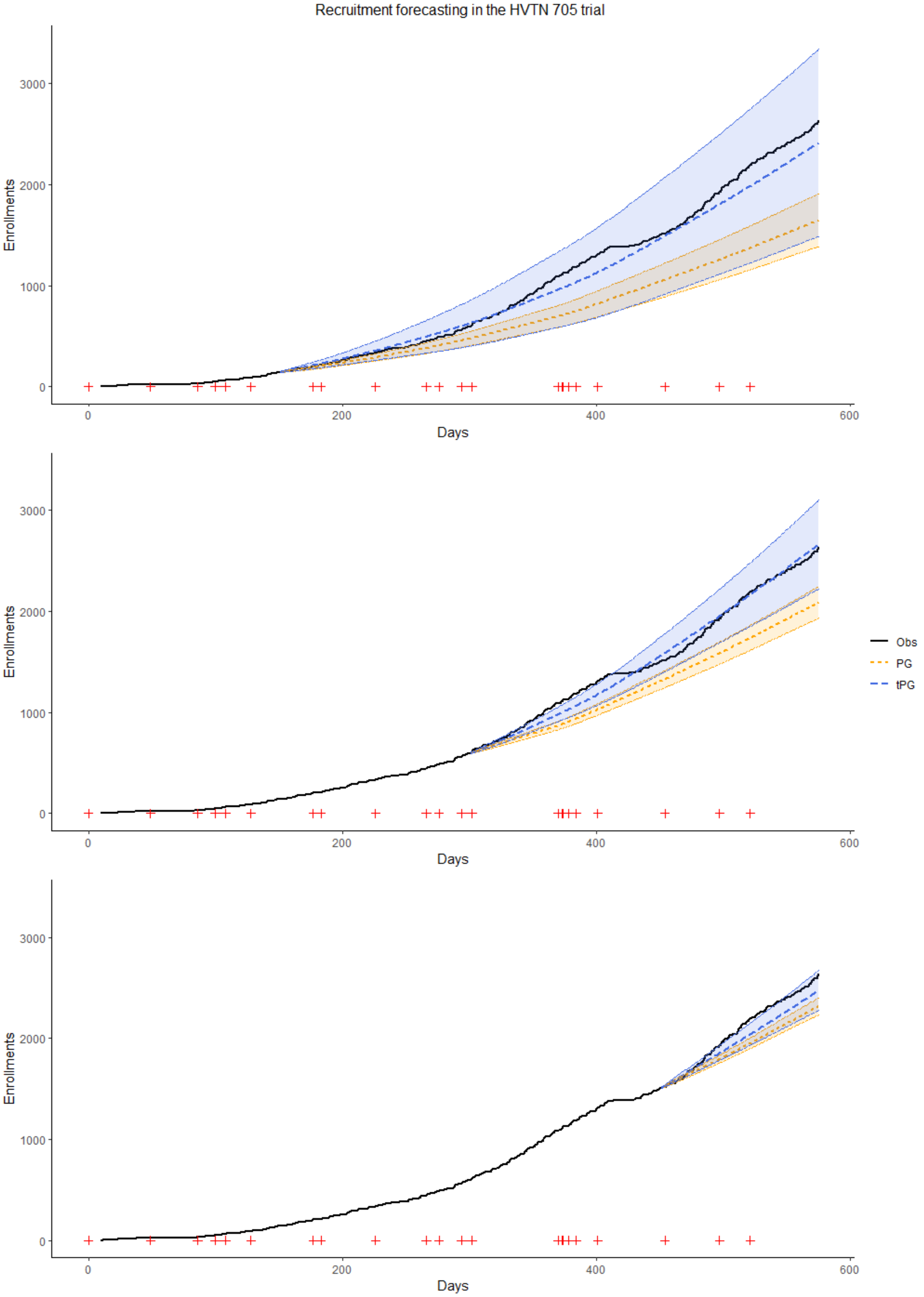

Figure 5.3: Enrollment predictions in the HVTN 705 study under the time-dependent and standard PG models compared to the observed recruitments at three interim times ($t_{int} = 150, 300,$ and 450 days). The shaded areas indicate the respective 95% credible intervals and the + marks represent the centers' initiation dates.

Similarly to the previous case study, the width of the credible intervals is vastly different under the two models. The CrIs estimated by the tPG model cover the observed recruitments over time and the time needed to achieve the target sample size. However, here the point estimates present significant differences as well. While in the forecastings of enrollments in the HVTN 702 trial the credible intervals under the tPG model entirely contain the CrIs estimated by the standard PG model (see Figure 5.2), in this case there is more separation, especially for the second interim time. More specifically, the forecasted recruitment process under the tPG model is remarkably close to the observed enrollments when t_{int} is equal to 150 and 300 days. This is quite an impressive result since by those interim times only 5% and 22% of the total study participants had already been recruited, and only 7 centers had been initiated after 150 days. Point estimates at the last interim time are slightly more off target, but the credible intervals still include the observed recruitment process.

Compared to HVTN 702, the model estimates are more stable as well. For both the first two interim times, the plateau point is estimated between 101 and 111 days following the center's initiation date. However, for $t_{int} = 450$, this point is estimated closer to the monitoring time (day 350).

5.4 Practical guide to the tPG R package

The analysis of the recruitment data from the HIV trials in the previous section was carried out via the R package tPG, which is available on GitHub. In this section, we show through simulated data the main functions embedded in the package which are needed to forecast enrollments and, consequently, estimate the remaining time to achieve a target sample size. The first step is the loading of the package:

```r
library(devtools)
install_github("aturchetta/tPG")
library(tPG)
set.seed(1)
```

Simulate data

Let us generate a synthetic data set. The function `sim_rec` allows the user to simulate recruitment data according to the generative process used in Turchetta et al. [2023] to validate the tPG model in a simulation study. For each center, the daily enrollments are sampled from a Poisson distribution with rate $\widetilde{\lambda}_i(t)$, where $\widetilde{\lambda}_i(t)$ is a draw from

$$\lambda_i(t) \sim \text{Gamma}\left(\alpha, \frac{\alpha}{f^q(t)}\right), \quad i = 1, \dots, C, \; t = 1, \dots, t_p$$

and $\widetilde{\lambda}_i(t) = \widetilde{\lambda}_i(t_p)$ for $t > t_p$. The mean parameter $f^q(t)$ representing the average recruitment rate over time can take the form of a scaled and shifted CDF or PDF of a Gamma random variable with shape parameter p_1 and rate p_2, i.e.,

$$f^1(x) = c_1 + c_2 \frac{p_2^{p_1}}{\Gamma(p_1)} e^{-p_2 x} x^{p_1 - 1},$$
$$f^2(x) = c_1 + c_2 \frac{p_2^{p_1}}{\Gamma(p_1)} \int_0^x e^{-p_2 u} u^{p_1 - 1} du.$$

Note that the choice of the Gamma distribution was made in light of its flexibility in generating different recruitment curves, and the use of its CDF function to simulate common recruitment behaviors has already been adopted in the literature for the validation of time-dependent recruitment models (e.g., Zhang and Long [2010] and Deng et al. [2017]). However, the mean parameter $f^q(t)$ is estimated via the B-spline model, hence this distribution is not to be confused with the Gamma distribution embedded into the tPG model.

The `sim_rec` function allows for the choice of all of the aforementioned simulation parameters. For the illustrative purposes of this example, we generate a data set similarly to the second recruitment scenario analyzed in Turchetta et al. [2023]: this entails a slow start in the recruitment intensity following the center's initiation date which then accelerates until the attainment of a plateau after 150 days. This recruitment progression is described by the CDF of the Gamma random variable (selected as `type = "cdf"` in the R function) with shape and rate parameters equal to 10 and 0.15 respectively, and the remaining parameters are $c_1 = 0.1$, $c_2 = 0.3$, $\alpha = 1$, and $t_p = 150$. We simulate data for 30 centers up to day 200 (`Tmax = 200`), which we can consider as the interim time of the analysis. The initiation dates can be either random or fixed. In the first case, the parameter `s_time` requires a vector of initiation times from which to sample the dates. The initiation time of the first center is fixed to 1. Otherwise, one can specify a vector of $C - 1$ fixed initiation times and select `random_init = F`. The following line of code simulates data according to the aforementioned parameters and random initiation times between day 1 and 100:

```
> sim_data <- sim_rec(alpha = 1, p1 = 10, p2 = 0.15, c2 = 0.3, c1 = 0.1,
type = "cdf", t_p = 150, Tmax = 200, s_time = 1:100, C_act = 30, random_init = T)
```

This function returns a list with two elements: the vector of initiation times and a list where each element contains the recruitment data of one center formatted as a data frame with two columns; the first column is the grid of recruitment days for the concerned center, and the second column indicates the respective number of enrollments on that day. Note that every recruitment day must be included even if no enrollments were registered on that day:

```
> sim_data$data[[1]][19:24,]
    t count
19 19     0
20 20     1
21 21     0
22 22     0
23 23     1
24 24     0
```

In our experience with the analysis of recruitment data from clinical studies, the data are often provided in two data sets: one containing the enrollment date for each participant along with the respective center and one containing the information on the recruitment sites which includes their activation date. Since the process of transitioning from one format to the other can be tedious, the supporting information includes an R code example where this process is laid out.

Model fitting

Once the data are formatted correctly, they can be fed to the tPG_est function. This function fits several B-spline models varying the polynomial degree and number of internal knots as well as their placement, and automatically selects the best-fitting model via the BIC. Currently, the function can accommodate only one internal knot between the center's initiation and the plateau point. Future versions of this package will allow for multiple knots. The user can choose a vector of candidate degree values through the parameter degree and knot placements via the parameter den as a function of the estimated plateau point, that is $\widehat{t_p}$/den. The likelihood function is optimized via the Bound Optimization BY Quadratic Approximation (BOBYQA) algorithm available in the nloptr package [Powell, 2009, Ypma et al., 2018]. Given the computational complexity, it is recommended to optimize the same likelihood multiple times using different initial values for the plateau point parameter t_p via the argument tp_start. For each combination of polynomial degree and knot placement, the tPG_est function will optimize the likelihood for each starting point and save the parameter estimates which result in the highest likelihood value. Setting the interim time at t_int=200, the following code estimates six models varying the polynomial degree (quadratic and cubic) and the internal knot placement (none, $\widehat{t_p}/2$, and $\widehat{t_p}/3$). The candidate initial values for t_p are 20, 60, 100, and 140 days.

,

```
> est <- tPG_est(data = sim_data$data, t_int = 200 , tp_start = c(20, 60, 100,
    140), degree = c(2, 3), den = c(NA, 2, 3) )
```

This function returns multiple objects, including the estimates of the parameters under each model and the best-fitting model selected via BIC:

,

```
> est$best_model
$`Chosen model`
[1] "degree: 2 , no internal knots"

$Parameters
        alpha      t_p  spline 1  spline 2   spline 3
[1,] 1.185418 139.3013 -2.233401 -3.988013 -0.6559678

$`Degree and knot`
[1]  2 NA
```

The above output indicates that the simplest model with second degree polynomial and no internal knot was selected. Additionally, the respective model estimates are returned, which indicate that the plateau point is estimated at 139 days.

Predictions

The tPG_pred function computes point estimates and credible intervals for future enrollments and the required time to achieve the targeted sample size. The function requires the output from the tPG_est function, the initiation dates of the centers through the argument start_init, and the future point in time or grid of time points for which predictions are computed through the tpred argument. The level of the credible interval can be set via the level argument. For example, if we are interested in the forecasted recruitments at day 500, the following line of code will deliver the estimate of the additional enrollments from the interim time (day 200) to day 500:

,

```
> pred_1 <- tPG_pred(est = est, start_init = sim_data$start, tpred = 500, level =
    0.95)
> pred$n_pred
         m     ci_l     ci_u   t
1 4575.471 3557.218 5593.723 500
```

If we are interested in the estimated time to achieve the desired sample size, we need to specify that through the argument N and input a grid of time points for tpred. For example, the estimated remaining time to recruit a total of 4000 participants is 253 days, with a 95% credible interval of [208, 325] days:

```
> pred_2 <- tPG_pred(est = est, start_init = sim_data$start,
+                  N=5000, tpred = 201:600, level = 0.95)
> pred_2$T_pred
[1] 253 208 325
```

In addition, the user can incorporate the opening of new centers by specifying their initiation date through the argument `start_add`. Say we want to evaluate the effect of the opening of 10 centers which will be initiated between day 255 and 300 every five days. Then the estimated remaining time decreases to 227 days (95% CrI: [195, 272]):

```
> pred_3 <- tPG_pred(est= est, start_init= sim_data$start, N=5000, tpred = 201:500,
+                  start_add = seq(255,300, length.out = 10) , level = 0.95)
> pred_3$T_pred
[1] 227 195 272
```

Finally, the results can be plotted via the `tPG_plot` function. The main arguments are `pred`, which requires the output from `tPG_pred`, and `print_T`, which controls whether the point estimate and CrIs for the time to achieve the sample size are printed. The remaining parameters are derivative of the standard `plot` function. Figure 5.4 displays the plots of the recruitment predictions with and without the opening of new centers generated via the following code:

```
par(mfrow=c(2,1))
tPG_plot(pred = pred_2, xlim=c(0,550), ylim=c(0,5500), print_T = T, xlab = "Days",
    ylab = "Enrollments",
         main = "Forecasted recruitments")
tPG_plot(pred = pred_3, xlim=c(0,550), ylim=c(0,5500), print_T = T, xlab = "Days",
    ylab = "Enrollments",
         main = "Forecasted recruitments with additional centers")
```

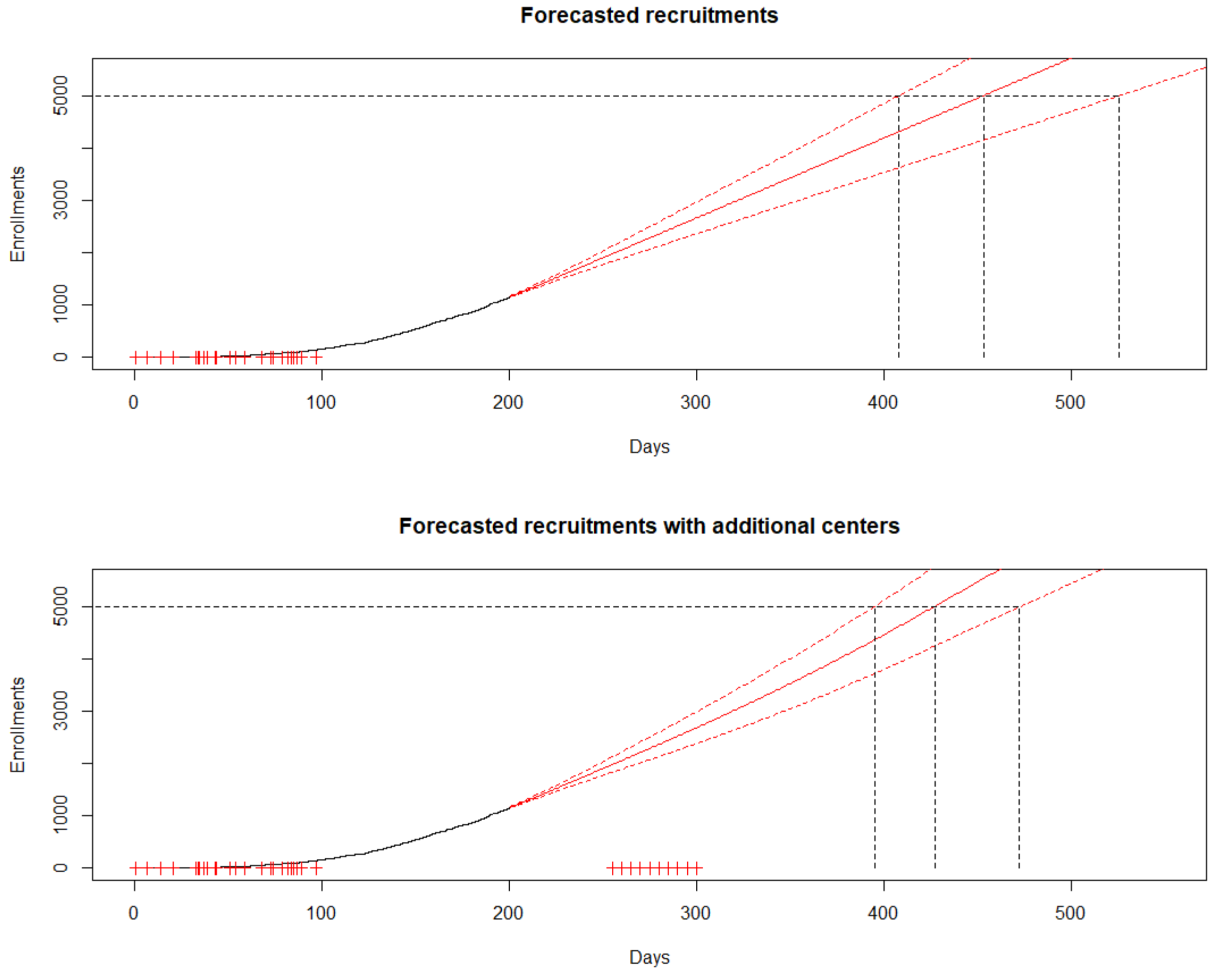

Figure 5.4: Plots generated by `tPG_plot`. The black solid line represents the observed recruitment up to the interim time and the red lines represent the point estimate (solid line) and credible intervals (dotted lines) of the predicted enrollments. The vertical lines indicate the point estimates and CrI limits of the time to achieve the target sample size (horizontal line). The + signs represent the centers' initiation date.

Inclusion of covariates

If the user wishes to include center-specific covariates, the functions `tPG_est_cov` and `tPG_pred_cov` can be used to estimate parameters and compute predictions. They require the same arguments as the regular estimating and prediction functions plus the `X` argument, which is a matrix where each column is a covariate.

5.5 Discussion

In this paper, we have validated the time-dependent Poisson-Gamma model on two recruitment data sets from randomized controlled trials and laid out a guide to its practical implementation in R. After outlining the model and its assumptions, we have forecasted the recruitment progression of two HIV vaccine trials conducted in Sub-Saharan Africa by the HIV Vaccine Trials Network and compared its performance to the standard Poisson-Gamma model.

The proposed model generally led to less biased point estimates and, crucially, to wider and more realistic credible intervals that contain the observed recruitment process, both in terms of predicted future enrollments and remaining time to achieve the target sample size. An analogous difference in CrIs' width was already observed in Zhang and Huang [2022] in the prediction of the recruitment process in two oncology trials, where the Poisson-Gamma model was compared to the Nonhomogenous Poisson Process (NHPP) model introduced in Zhang and Long [2010], another time-dependent recruitment model. In those trials, the PG model did achieve a good performance, and even outperformed the more complex NHPP model. The authors argued that this is a reasonable outcome in oncology trials, where patients arrive to the recruitment centers at a generally stable rate. We observed a similar pattern in The Promotion of Breastfeeding Intervention Trial (PROBIT), a randomized clinical trial where the participants were mother-infant pairs [Kramer et al., 2001]. Since the effective date of enrollment was the birth of the newborn, the recruitments were stable over time. The tPG model estimated the plateau point close to the initiation of the center, hence its predictions were very close to the ones of the standard PG model, although both methodologies failed to accurately capture the recruitment process. The results are presented in the appendix. However, this constant-intensity assumption was not met in the two HIV vaccine trials analyzed in this paper, where the PG model led to biased and overly narrow credible intervals that did not contain the observed recruitment process, giving a false sense of precision. On the other hand, while the tPG model led to a positive performance in terms of forecastings in both trials, some limitations need to be discussed. The assumption that by the interim time enough centers have passed the stabilization point to allow for its estimation might not be met. A warning sign that this might be the case is when the plateau point is estimated close to the monitoring time. Additionally, it is possible that perturbations in the recruitment progression following the interim time lead to a different estimated stabilization point when more data are accrued, or the selection of a different B-spline model. This aspect may be exacerbated by the presence of breaks in recruitment. While it has been proven in Minois et al. [2017b] that systematic breaks are not an issue in the standard PG model, they represent a hurdle in the time-dependent model, as they it make more difficult for the model to accurately estimate the stabilization point. This might partially explain why the estimated plateau point moved forward in time along with the monitoring time of the analysis in HVTN 702 and in the last monitoring time of HVTN 705, which came soon after the December holiday period. We made an adjustment in HVTN 702 by removing some days in the December break from the data set, however this is a simple solution that does not account for possible upticks in recruitment before and after the break, and it was not applicable to HVTN 705 since some centers did not stop recruiting. A formal inclusion of systematic breaks or slowdowns in recruitment into the time-dependent PG model is an interesting avenue for future research. However, it should be noted that this drawback in the current model formulation does not necessarily lead to more biased estimates, but rather to larger credible intervals since fewer data points are incorporated into the posterior distribution of the recruitment rates.

Finally, we provided a tutorial on the newly developed `tPG` R package, which implements the necessary functions for the estimation, prediction, and basic visualization of the recruitment process via the time-dependent Poisson-Gamma model. There are some improvements that need to be implemented in the package, such as allowing for more than one internal knot in the B-spline model and including new functions to ease the initial data manipulation step. Hopefully, this paper will serve as a step forward in facilitating the adoption and implementation of recruitment modeling in practice.

Acknowledgments

AT is supported by a doctoral fellowship from the Fonds de Recherche du Québec (FRQ) - Nature et Technologie and by funds from a Recherche en Équipe grant from the FRQ – Nature et Technologie awarded to DAS. DAS and EEMM acknowledge support from Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). EEMM is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine and acknowledges the support of a chercheur de mérite career award from the FRQ - Santé. The HVTN 702 and HVTN 705 clinical trials were conducted by the HIV Vaccine Trial Network. HVTN 702 was co-funded by the National Institute of Allergy and Infectious Disease, Bill & Melinda Gates Foundation, GSK, and Sanofi Pasteur. HVTN 705 was co-funded by Janssen Vaccines & Prevention BV; Bill & Melinda Gates Foundation; National Institute of Allergy and Infectious Diseases/Division of AIDS; Ragon Institute of the Massachusetts Institute of Technology, Massachusetts General Hospital, and Harvard; and US Army Medical Materiel Development Activity.

Disclaimer

The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of any of the aforementioned funding agencies.

Chapter 6

Conclusion

6.1 Summary

In this thesis, I have illustrated the development of Bayesian approaches that can aid the trialist in challenging aspects of the planning and monitoring phases of clinical studies.

In the first manuscript presented in Chapter 3, I have highlighted some difficulties in determining the appropriate sample size for a SMART study which guarantees an adequate level of power, especially in relation to the uncertainty and plausibility around the assumptions and specification of key design parameters at the pre-experimental stage. Building on the work of Oetting et al. [2011], by leveraging data from pilot trials, I developed a flexible and robust sample size methodology within the Bayesian framework for sizing SMART studies in order to compare two adaptive treatment strategies that start with a different intervention. With respect to the standard frequentist calculations introduced in Oetting et al. [2011], the proposed methodology has numerous advantages. The implementation of the 'two priors' approach in the sample size determination allows researchers to incorporate pre-trial knowledge via the analysis prior and, most importantly, uncertainty around the minimal detectable difference via the design prior. This formulation leads to a methodology that is not affected by the local optimality issue, which is a common problem in frequentist sample size determination where a single value for the effect of treatment (or, in this case, of the treatment strategy) under the alternative hypothesis must be posited to power the study. Even slight deviations from this quantity may lead to a substantial decrease in the power of the study. This issue is particularly problematic when the information on the effects of the treatments being compared is limited. In the case of SMARTs, this problem is exacerbated, as the information regarding treatment strategies from previous studies is even more challenging to obtain. Additionally, with respect to standard RCTs, the sample size determination for SMARTs necessitates the elicitation of an additional parameter, which is the probability of response to the initial treatments. It has been argued that the elicitation of these design parameters can be informed by a pilot SMART [Almirall et al., 2012]. However, pilot SMARTs are generally not sized with the purpose of delivering precise estimates of these quantities, which, given the numerous treatment paths a participant can take in a SMART, might be based on a very small number of subjects. A precision-based approach for sizing pilot SMARTs in order to confine the estimated marginal mean outcome of a treatment strategy within a pre-specified margin of error has been proposed [Yan et al., 2021]. However, while this approach can lead to a more informed decision regarding the difference between the effects of two treatment strategies, the local optimality issue remains, and the approach does not address the precision of the estimates of the strategies' standard deviation needed to calculate the standardized effect size required to size a full-scale SMART. By eliciting a prior distribution on the minimal detectable difference, the approach that I put forward in Chapter 3 directly takes into account the uncertainty around the effects of the treatment strategies being compared, leading to a more robust methodology. Additionally, the information gained from the pilot study or external sources can be incorporated into the analysis stage via the analysis prior.

Another important advantage of the proposed methodology is that it does not rely on the assumptions regarding the conditional variance of the outcome and the intermediate response rates to initial treatments. In fact, in the frequentist sample size calculations, it is assumed that the variability of the outcome around the strategy mean for both responders and non-responders to the first intervention is not greater than the variance of the strategy mean. This assumption is quite abstract, but it is needed to avoid the specification of the variance of the estimated strategy means, as it leads to a sample size formula that only depends on the standardized effect size and response rates to the initial treatments, which are assumed to be equal. In practice, I found that this assumption was rarely exactly met for both responders and non-responders, and the greater the difference in variability between the two groups, the more the departure from the assumption weighed on the performance of the frequentist sample size estimates. In the proposed Bayesian approach, we take a different route by leveraging data from the pilot study. The use of crude estimates from pilot studies of the variance components for the purpose of sizing a full-scale trial has been widely criticized, as the estimates can be unreliable and lead to underpowered studies. To avoid this drawback and take into account the uncertainty around these estimates, the approach put forward in the first manuscript entails the estimation of a posterior distribution of the variance components based on the pilot data. This distribution is then used to marginalize the Bayesian power function, achieving a marginal power function that does not depend on either the variance of the strategy means or the probability of response to initial treatments. I have examined the performance of this approach compared to the standard frequentist methodology in a simulation study that included different scenarios and a wide range of degrees of model misspecification. At the expense of a generally higher sample size, the results demonstrate that the proposed approach was more robust to deviations from the posited standardized mean difference/minimal detectable difference, and it was not affected by the violation of the frequentist assumption on the conditional variance of the outcome. Moreover, while the approach is designed for continuous outcomes, the simulation study showed that the Normal approximation for binary outcomes led to the same level of performance in the scenarios we considered. We implemented our approach to estimate the sample size of a full-scale

SMART, namely the Internet-Based Adaptive Stress Management SMART, using data from its pilot study conducted in two Canadian provinces. Finally, the supplementary material illustrates how to implement the proposed sample size approach via the `bayesSMARTsize` R package accessible on GitHub.

In the second and third manuscripts, I focused on the monitoring phase of multicenter clinical studies, developing a Bayesian approach to forecast the recruitments of an ongoing study. In Section 2.4, I reviewed the approaches introduced for this task and their limitations. In the second manuscript presented in Chapter 4, I introduced a novel methodology for forecasting recruitments as a time-dependent extension of the Poisson-Gamma model developed by Anisimov and Fedorov [2007a, 2007b]. In fact, although the PG method has been validated on several data sets from real clinical trials, its assumption that the centers' enrollment rates remain constant throughout the recruitment process is often not met in practice. The proposed time-dependent Poisson-Gamma model alleviates this restriction by allowing the rates to vary with time until the attainment of a plateau at an unknown time point. More specifically, the centers' recruitment rates are assumed to originate from a Gamma distribution whose log-scaled mean parameter is modelled via B-splines. Since the proposed model is an empirical Bayes method, the model's parameters are estimated via the likelihood function using the recruitment data collected from the ongoing study and plugged into the distribution of the recruitment rates. The model was first assessed in a simulation study for a wide range of recruitment scenarios comparing its predictive performance to the standard PG model. In settings where the recruitment intensity was assumed to be time-dependent, it is unsurprising that the proposed approach outperformed the PG model. However, in the setting where recruitment rates were constant over time, the two approaches led to a comparable performance, suggesting that the proposed recruitment model can be framed as a time-dependent generalization of the standard PG model. Finally, both models were implemented in a case study to forecast recruitments in the Canadian Co-infection Cohort study, an ongoing prospective observational study that encompasses 19 recruitment sites across Canada. In this study, the centers generally experienced a fast start in recruitments, which slowed down with time. Since the standard Poisson-Gamma model does not take into account this inhomogeneity, the resulting recruitment predictions were upwardly biased. On the other hand, the proposed time-dependent model was able to adjust for this drawback by estimating the rates to be constant only after 110 weeks following the centers' opening, delivering significantly more accurate forecastings.

In the third manuscript presented in Chapter 5, the time-dependent Poisson-Gamma model was further validated on the recruitment data from two HIV clinical trials to show the applicability of the proposed approach in the context of randomized controlled trials. The two trials were conducted in multiple Sub-Saharan African countries, and they both sought to evaluate the safety and efficacy of two vaccine candidates in preventing HIV infections. I demonstrated that the time-dependent approach delivered accurate predictions outperforming the standard Poisson-Gamma model. In addition to the difference in point estimates, these two case studies highlighted an important difference between the two approaches in terms of the width of the resulting credible intervals. By assuming constant recruitment rates over time, the standard Poisson-Gamma model includes the entirety of the accrued enrollment data in the posterior distribution of the rates. However, if this assumption is not met, not only the point estimates are biased, but the credible intervals are overly narrow. On the other hand, the time-dependent Poisson-Gamma model incorporates into the distribution of the rates only the recruitment data collected after the stabilization point, whereas the previous enrollments are solely used to estimate the model's parameters. This approach generates credible intervals that are generally wider and more realistic. This difference between approaches is evident in both of the case studies analyzed in this manuscript.

A secondary aim of the manuscript was to illustrate a step-by-step guide on how to implement the proposed model to facilitate its adoption by statisticians and investigators. In fact, while the body of literature on statistical methods for forecasting enrollments in clinical studies has been growing in recent years, these approaches are rarely implemented in practice. Therefore, Chapter 5 includes a tutorial on the R package `tPG`, which includes the necessary functions for the estimation, prediction, and basic visualization of the recruitment process via the time-dependent Poisson-Gamma model.

6.2 Limitations and future work

Some limitations of the methodologies developed in this thesis need to be discussed.

The sample size approach for SMART studies introduced in the first manuscript relies on the availability of pilot data to estimate the posterior distribution of the relevant variance components. While full-scale SMARTs are often preceded by a pilot study to assess the feasibility of the study, this might not always be the case. In the discussion section of Chapter 3, I proposed a simulation-based algorithm to overcome the lack of pilot data, but further work is needed to assess its plausibility, as it requires the specification of several parameters to generate the simulated data. Furthermore, the use of pilot data to estimate the posterior distribution of the variance components in order to marginalize the power function for the full-scale trial implies that the sample size is a random variable. The simulation study of this manuscript showed that the variability in sample size estimates across data replications can be significant. The main driver of this variability is the small size of the pilot studies. A precision-based approach for sizing pilot SMARTs was introduced in Yan et al. [2021]. Although this approach targets the precision of the estimated mean outcome of a treatment strategy, it would be an interesting avenue for future research to develop an approach that targets the accuracy of the variance of the estimated strategy mean, which would ensure a higher degree of consistency in the sample size determination of the full-scale trial.

Furthermore, the simulation and case studies illustrated in the first manuscript showed the importance of the choice of hyperparameters for the design and analysis prior distributions, as the estimated sample size varied significantly across the different choices. While this level of flexibility is an advantage with respect to the frequentist methodology, a certain degree of subjectivity is often needed to select these hyperparameters. In the manuscript, I presented the sensitivity of the sample size estimate across several combinations of hyperparameters, however their selection was indeed subjective. The Bayesian literature offers several approaches to select appropriate parameters, especially in regard to the analysis prior since it can formally incorporate knowledge from pilot data or external studies. It can be attractive to attribute a large weight to this pre-trial information by setting a low value for the variance of the analysis prior, as this could potentially result in a significant reduction in sample size. However, one must ensure that this choice is not overly influential on the final results of the study. A more rigorous approach to incorporate pre-trial knowledge in the context of SMART studies could be further explored.

The proposed sample size methodology relies on the postulation of a minimal detectable difference between treatment strategies instead of the standardized effect size. Although this can be an advantage in terms of interpretability, it is also a limitation, as there are instances where the standardized effect size might be preferable.

Finally, our approach targeted the comparison of two strategies that start with a different intervention in a two-stage SMART design where responders to the first treatment are not re-randomized. As discussed in the literature review in Chapter 2, this is the most common SMART design. However, in future work, the methodology could be extended in various directions varying the number of stages and treatments available to responders and allowing for the comparison of strategies which start with the same intervention, which would require an adjustment for the correlation between strategies. Furthermore, this approach was developed for SMARTs with a continuous final outcome. The simulation study showed that the Normal approximation for binary outcomes is satisfactory, however, the ad-hoc extension of this approach to binary and other types of outcomes is a possible avenue for further development.

I now turn to some limitations regarding the time-dependent Poisson-Gamma model introduced in the second manuscript. This approach targets the specific case where recruitment rates vary with time until a stabilization point, and the recruitment intensity is constant thereafter. This assumption held true in the three case studies presented in Chapters 4 and 5, and it has also been proposed in other publications. However, it might be unrealistic for other studies. In the supplementary material of the third manuscript (Appendix C), I presented a counterexample, the PROBIT study, where the proposed method did not deliver adequate predictions. More specifically, the recruitment rates appeared to be constant, which led the time-dependent approach to estimate the plateau point close to the center's opening, yielding nearly the same predictions as those produced under the standard PG model. However, both approaches failed to accurately capture the recruitment process. In Chapter 2, I reviewed several time-dependent models that entail different assumptions regarding the recruitment progression over time. Since the forecasting of recruitments is essentially an extrapolation problem, the method to be implemented should depend on the recruitment expectations or historical data. In addition, it would be useful to implement multiple approaches and assess the sensitivity in forecastings under the different underlying assumptions.

Moreover, the proposed method relies on the availability of recruitment data past the stabilization point for an adequate number of centers. The simulation study in the second manuscript highlighted that the model struggled to achieve good predictive properties for the more elaborate recruitment curves when a sizable number of centers has not passed this point. The propensity of the approach to select more complex B-spline models via the BIC in these scenarios suggests that increasing the number of candidate models by varying the number and placement of internal knots might lead to a better performance. However, this should be assessed in further simulation studies. The assumption itself that by the interim time when the predictions are computed at least one center has passed the plateau point might not be met. In this case, the plateau point is likely to be estimated close to the interim time, and the resulting forecastings are produced under the assumption that the recruitment rates will be constant from that point onward. Additionally, the plateau point is assumed to be the same across centers, which is an assumption that could be relaxed in future work by assuming its dependence on center-specific covariates. It is also important to note that the reliance on the recruitment data from the ongoing study makes the approach unusable for predicting enrollments in the planning stage of the study. It would be an interesting project to extend this approach to the pre-trial phase by leveraging historical data similarly to the approach developed by Minois et al. [2017a] for the standard Poisson-Gamma model.

Finally, the two case studies analyzed in Chapter 5 highlighted the issue of breaks during the recruitment period. Minois et al. [2017b] analyzed this problem in the context of constant recruitment rates, concluding that systematic breaks in recruitment did not affect the accuracy of the standard PG model. However, in time-dependent models, this is a more relevant problem, as the breaks influence the estimation of the stabilization point. In the manuscript, I have adjusted the model for the case study where the recruitment process was completely halted during the December holiday period by removing these days from the data set. This is a simplistic solution, and the inclusion of systematic recruitment breaks or slowdowns into the proposed approach could be addressed in future research.

6.3 Concluding remarks

This thesis focused on the development of Bayesian tools which address challenges in the design and monitoring phases of clinical trials. I have introduced a Bayesian approach that relies on minimal assumptions to determine the sample size of SMART studies and developed a flexible approach to monitor recruitments in multicenter trials. Both methodologies were assessed in extensive simulation studies and validated in multiple case studies, which showed their practical applicability. Additionally, to facilitate their adoption, this thesis includes step-by-step tutorials on the R packages developed to implement these approaches.

APPENDIX

APPENDIX A

Appendix to Manuscript 1

Web Appendix A: steps for inequlity 3.4

Since

$$\theta|V_n \sim \mathcal{N}\left(\frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2},\ \left(\frac{1}{\sigma_0^2} + \frac{n}{\tau^2}\right)^{-1}\right),$$

it follows that

$$Z = \left(\theta - \frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2}\right)\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\tau^2}}\Big|V_n \sim \mathcal{N}(0, 1).$$

Therefore,

$$\begin{aligned}
&\Pr_{\pi(\cdot|V_n)}(\theta > 0) \geq 1-\epsilon \\
\iff &\Pr_{\pi(\cdot|V_n)}\left(Z > -\frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2}\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{\tau^2}}\right) \geq 1-\epsilon \\
\iff &1-\Phi\left(-\frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2}\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{\tau^2}}\right) \geq 1-\epsilon \\
\iff &\Phi\left(-\frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2}\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{\tau^2}}\right) \leq \epsilon \\
\iff & -\frac{\tau^2\theta_0 + n\sigma_0^2 V_n}{\tau^2 + n\sigma_0^2}\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{\tau^2}} \leq z_\epsilon \\
\iff &V_n \geq -\frac{z_\epsilon \tau\sqrt{\tau^2 + n\sigma_0}}{n\sigma_0} - \frac{\theta_0\tau}{n\sigma_0^2}.
\end{aligned}$$

Web Appendix B: data generating mechanism

In this section, we outline the data generating mechanism that was employed for the simulation study analyzed in the paper. To mimic the common practice in SMART studies of dichotomizing continuous outcomes to identify responders and non-responders, binary outcomes are generated by dichotomizing a draw from a Normal distribution using the appropriate quantile z so as to achieve the expected response rates. These continuous latent outcomes are indicated with R^* and Y^*, where the former denotes the sample from the Normal distribution employed to generate the intermediate outcome R, whereas the latter is used to generate the final outcome Y in the binary outcome scenarios. Finally, $p^*_{a_1}$ and $p^*_{a_1,a_2}$ indicate the misspecified response rates to the first and second stage intervention respectively – the latter of which is only relevant in the binary outcome scenarios – and TN indicates the truncated Normal random variable where the four parameters are mean, variance, and lower and upper truncation limits. The data generating algorithm is outlined below.

Algorithm 1: Data generating mechanism

```
$p^*_{a_1} \sim TN(p_{a_1}, \sigma^2, 0, 1)$
if $Y$ is binary then
    $p^*_{a_1,a_2} \sim TN(p_{a_1,a_2}, \sigma^2, 0, 1)$

for $i = 1, \ldots, n$ do
    1) $A_1 = a_1 \sim \text{Bern}(0.5)$
    2) $R^* \sim \mathcal{N}(0, 1)$
    3) if $R^* < z_{p^*_{a_1}}$ then
        $R = 1$
        $\Pr(A_2 = a_1 | R = 1) = 1$
    else
        $R = 0$
        $A_2 = a_2 | R = 0 \sim \text{Bern}(0.5)$
    4) if $Y$ is continuous then
        $Y | A_1, R, A_2 \sim \mathcal{N}\left(\text{E}[Y | A_1, R, A_2], \text{Var}[Y | A_1, R, A_2]\right)$
    else if $Y$ is binary then
        if $R$=0 then
            $Y^* \sim \mathcal{N}(0, 1)$
            if $Y^* < z_{p^*_{a_1,a_2}}$ then
                $Y = 1$
            else
                $Y = 0$
        else
            $Y = 1$
```

Web Appendix C: simulation study results

In this supplementary section we showcase the full results of the simulation study. Table A.1 depicts the performance of the frequentist formula in terms of power in the binary outcome case (Scenarios 3 and 4) and Table A.2 shows the sample size estimates under the frequentist approach. Tables A.3 and A.4 outline the properties of the proposed Bayesian sample size methodology in Scenario 2 when the analysis prior mean θ_0 is set to 0 and to the difference between strategy means estimated via pilot data $\widehat{\theta}^p$ respectively. Tables A.5-A.6 and A.7-A.8 provide the same information for Scenarios 3 and 4. Finally, Tables A.9 and A.10 illustrate the simulated type I error under the Bayesian methodology in the continuous and binary case respectively. The parameters chosen for this simulation study were informed by real data and previous works. Specifically, we considered effect sizes and probabilities of response to initial treatments similar to those used in Scott et al. (2007), while the parameters for the sizing of pilot studies were informed by two pilot SMARTs, one of which is the Internet-Based Adaptive Stress Management Pilot SMART outlined in Section 3.4. During the simulation study phase of the drafting of this paper, we tested a variety of scenarios that included lower levels of the effect size and the response rate to the initial treatment in both the binary and continuous outcome cases. In general, we found that the conclusions were similar, and the four scenarios that were further investigated and presented in this paper are those that clearly exemplify the differences in performance between the Bayesian and frequentist methodologies.

Table A.1: Simulated power under the frequentist calculations in the binary outcome scenarios, i.e. Scenarios 3 (left) and 4 (right), for different degrees of model misspecification.

		Scenario 3					Scenario 4		
		Response SD							
		0	0.01	0.02	0.03	0.04	0	0.01	0.02
% bias of δ	0	0.89	0.88	0.85	0.84	0.80	0.83	0.82	0.79
	5	0.87	0.85	0.84	0.80	0.77	0.81	0.80	0.77
	10	0.85	0.83	0.81	0.79	0.74	0.78	0.75	0.74
	15	0.82	0.80	0.80	0.76	0.74	0.74	0.71	0.72
	20	0.78	0.77	0.75	0.75	0.72	0.70	0.71	0.69
	25	0.74	0.74	0.74	0.72	0.70	0.67	0.68	0.66

Table A.2: Frequentist sample size estimates for each scenario and different degrees of overestimation of the standardized effect size.

		Scenario			
		1	2	3	4
% bias of δ	0	302	628	728	1516
	5	274	570	660	1376
	10	250	520	602	1254
	15	228	476	550	1148
	20	210	436	506	1054
	25	194	402	466	972

Table A.3: Power and average sample size (first and third quartile in brackets) under Scenario 2 (continuous outcome) computed using the Bayesian ‘two priors’ approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to 0.

		σ_d							
		0		0.2		0.5		0.8	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n
1	100	0.88	741 (617, 856)	0.87	769 (646, 880)	0.93	942 (785, 1085)	0.98	1452 (1221, 1671)
	3	0.88	767 (648, 882)	0.88	785 (658, 907)	0.92	963 (804, 1109)	0.98	1493 (1249, 1726)
	2	0.87	782 (651, 905)	0.88	810 (676, 933)	0.93	990 (828, 1134)	0.98	1516 (1269, 1750)
	1	0.86	899 (754, 1031)	0.88	933 (780, 1077)	0.92	1119 (935, 1291)	0.98	1674 (1410, 1930)
2	100	0.80	736 (615, 851)	0.80	764 (638, 881)	0.85	946 (791, 1089)	0.90	1444 (1204, 1673)
	3	0.79	760 (630, 878)	0.80	781 (655, 900)	0.83	964 (806, 1118)	0.90	1469 (1223, 1699)
	2	0.78	780 (647, 901)	0.80	816 (682, 936)	0.84	992 (836, 1138)	0.91	1500 (1247, 1735)
	1	0.77	898 (752, 1027)	0.78	928 (778, 1069)	0.84	1126 (945, 1301)	0.90	1665 (1391, 1923)
3	100	0.74	473 (397, 544)	0.73	484 (405, 556)	0.80	551 (460, 637)	0.85	713 (597, 822)
	3	0.74	496 (419, 568)	0.73	506 (429, 580)	0.79	577 (484, 662)	0.87	746 (624, 863)
	2	0.73	520 (437, 598)	0.72	531 (449, 609)	0.79	599 (502, 690)	0.86	768 (643, 883)
	1	0.70	618 (519, 715)	0.70	636 (535, 735)	0.77	717 (598, 824)	0.86	904 (761, 1039)
4	100	0.68	471 (391, 546)	0.68	484 (401, 558)	0.72	547 (456, 635)	0.78	709 (587, 823)
	3	0.67	493 (412, 567)	0.68	504 (421, 580)	0.72	571 (478, 664)	0.79	740 (621, 853)
	2	0.68	518 (434, 594)	0.68	528 (441, 610)	0.70	595 (498, 684)	0.79	768 (641, 886)
	1	0.65	616 (515, 710)	0.65	633 (526, 731)	0.69	707 (592, 815)	0.77	890 (754, 1024)

Table A.4: Power and average sample size (first and third quartile in brackets) under Scenario 2 (continuous outcome) computed using the Bayesian ‘two priors’ approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to $\widehat{\mu}^p_{\bar{d}^1_k} - \widehat{\mu}^p_{\bar{d}^2_k}$ (estimated from the simulated pilot study).

		σ_d							
		0		0.2		0.5		0.8	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n
1	100	0.88	739 (618, 853)	0.87	765 (640, 883)	0.93	942 (789, 1086)	0.98	1459 (1224, 1673)
	5	0.88	723 (596, 837)	0.88	746 (615, 870)	0.92	919 (759, 1068)	0.98	1419 (1187, 1637)
	4	0.87	716 (582, 843)	0.88	737 (600, 861)	0.93	904 (742, 1056)	0.98	1402 (1159, 1635)
	3	0.86	685 (552, 814)	0.88	716 (570, 846)	0.92	876 (693, 1043)	0.98	1353 (1087, 1593)
2	100	0.80	739 (623, 848)	0.80	764 (635, 882)	0.83	941 (782, 1088)	0.91	1446 (1203, 1668)
	5	0.80	720 (590, 844)	0.78	749 (616, 875)	0.84	917 (752, 1070)	0.90	1411 (1159, 1650)
	4	0.79	698 (563, 827)	0.80	726 (584, 856)	0.84	904 (726, 1069)	0.90	1398 (1127, 1651)
	3	0.80	683 (539, 817)	0.80	709 (563, 842)	0.85	866 (678, 1032)	0.90	1344 (1050, 1603)
3	100	0.74	471 (396, 543)	0.76	487 (405, 561)	0.79	553 (461, 637)	0.87	714 (602, 820)
	5	0.73	461 (378, 539)	0.75	470 (384, 549)	0.79	536 (439, 621)	0.87	695 (572, 805)
	4	0.74	455 (368, 535)	0.76	466 (378, 543)	0.80	525 (428, 612)	0.86	676 (548, 794)
	3	0.75	431 (337, 519)	0.76	449 (353, 534)	0.79	507 (399, 602)	0.87	656 (513, 785)
4	100	0.70	474 (398, 546)	0.71	483 (404, 554)	0.71	545 (455, 632)	0.80	714 (598, 823)
	5	0.69	456 (374, 533)	0.71	469 (382, 550)	0.72	531 (432, 623)	0.78	685 (562, 801)
	4	0.70	450 (363, 530)	0.72	463 (374, 544)	0.74	522 (416, 617)	0.79	678 (545, 800)
	3	0.71	434 (332, 522)	0.72	447 (341, 540)	0.74	509 (390, 617)	0.78	659 (510, 787)

Table A.5: Power and average sample size (first and third quartile in brackets) under Scenario 3 (binary outcome) computed using the Bayesian ‘two priors’ approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to 0.

		σ_d											
		0		0.01		0.02		0.03		0.04		0.05	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n	Power	n	Power	n
1	5	0.88	719 (652, 788)	0.90	730 (661, 796)	0.90	776 (700, 853)	0.93	863 (786, 943)	0.96	997 (909, 1094)	0.98	1215 (1105, 1329)
	0.5	0.89	724 (657, 795)	0.90	740 (672, 809)	0.90	784 (709, 860)	0.93	864 (785, 943)	0.96	995 (903, 1091)	0.98	1219 (1107, 1339)
	0.3	0.89	735 (664, 806)	0.90	747 (678, 817)	0.90	788 (711, 865)	0.93	871 (787, 955)	0.96	1007 (911, 1106)	0.98	1229 (1114, 1346)
	0.1	0.89	810 (732, 889)	0.90	822 (748, 900)	0.91	868 (787, 947)	0.93	955 (869, 1044)	0.96	1097 (992, 1204)	0.98	1317 (1191, 1446)
2	5	0.79	717 (647, 787)	0.78	729 (658, 803)	0.79	772 (696, 851)	0.83	853 (770, 939)	0.84	997 (898, 1099)	0.88	1204 (1094, 1323)
	0.5	0.79	719 (654, 785)	0.79	734 (662, 808)	0.81	779 (700, 857)	0.83	858 (780, 941)	0.85	995 (900, 1096)	0.88	1207 (1094, 1321)
	0.3	0.79	729 (658, 801)	0.78	739 (666, 813)	0.81	784 (708, 861)	0.81	871 (789, 960)	0.85	1002 (908, 1102)	0.88	1222 (1110, 1341)
	0.1	0.78	803 (728, 881)	0.78	816 (737, 898)	0.79	859 (775, 941)	0.82	951 (862, 1046)	0.83	1087 (980, 1198)	0.88	1316 (1189, 1441)
3	5	0.75	462 (417, 505)	0.75	466 (424, 512)	0.76	483 (438, 528)	0.78	517 (470, 566)	0.82	565 (511, 619)	0.86	634 (574, 695)
	0.5	0.74	465 (420, 509)	0.75	471 (427, 516)	0.76	490 (443, 537)	0.78	519 (469, 567)	0.82	570 (516, 624)	0.86	639 (577, 702)
	0.3	0.73	472 (426, 517)	0.75	477 (433, 522)	0.76	496 (450, 543)	0.78	527 (478, 577)	0.83	577 (525, 630)	0.86	652 (592, 714)
	0.1	0.72	541 (491, 591)	0.73	549 (500, 600)	0.76	570 (520, 622)	0.79	602 (546, 656)	0.80	654 (592, 716)	0.86	730 (662, 800)
4	5	0.69	458 (415, 503)	0.68	466 (422, 510)	0.70	480 (435, 527)	0.71	515 (467, 567)	0.74	561 (507, 616)	0.75	631 (568, 696)
	0.5	0.69	464 (420, 509)	0.69	468 (423, 513)	0.70	485 (442, 530)	0.71	520 (471, 569)	0.73	565 (509, 623)	0.78	637 (576, 701)
	0.3	0.69	469 (424, 515)	0.67	475 (429, 523)	0.70	492 (445, 540)	0.71	526 (476, 577)	0.73	572 (517, 628)	0.76	646 (585, 705)
	0.1	0.66	540 (487, 592)	0.68	546 (493, 599)	0.67	564 (510, 621)	0.69	597 (545, 653)	0.74	655 (590, 722)	0.75	726 (657, 797)

Table A.6: Power and average sample size (first and third quartile in brackets) under Scenario 3 (binary outcome) computed using the Bayesian 'two priors' approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to $\widehat{\mu}^p_{\bar{d}^1_k} - \widehat{\mu}^p_{\bar{d}^2_k}$ (estimated from the simulated pilot study).

		σ_d											
		0		0.01		0.02		0.03		0.04		0.05	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n	Power	n	Power	n
1	5	0.89	719 (650, 790)	0.90	734 (665, 804)	0.91	777 (705, 852)	0.93	862 (784, 945)	0.96	996 (903, 1089)	0.98	1218 (1105, 1334)
	0.5	0.89	711 (643, 784)	0.89	722 (651, 799)	0.91	770 (694, 849)	0.93	850 (770, 935)	0.96	984 (889, 1086)	0.98	1199 (1079, 1321)
	0.3	0.89	697 (621, 780)	0.90	709 (632, 791)	0.92	752 (665, 839)	0.93	824 (732, 920)	0.95	962 (855, 1072)	0.98	1175 (1047, 1310)
	0.2	0.89	661 (555, 767)	0.90	678 (577, 786)	0.91	721 (606, 841)	0.93	786 (668, 913)	0.96	908 (763, 1060)	0.98	1114 (945, 1287)
2	5	0.78	718 (651, 787)	0.80	731 (660, 803)	0.79	773 (696, 849)	0.81	853 (772, 934)	0.84	985 (890, 1080)	0.88	1208 (1092, 1327)
	0.5	0.79	709 (638, 781)	0.79	717 (643, 791)	0.80	763 (688, 840)	0.82	847 (765, 935)	0.84	980 (887, 1080)	0.87	1192 (1071, 1315)
	0.3	0.79	690 (612, 774)	0.79	706 (628, 789)	0.81	747 (665, 838)	0.82	822 (731, 918)	0.84	955 (842, 1073)	0.87	1164 (1036, 1302)
	0.2	0.79	657 (549, 766)	0.79	671 (559, 787)	0.80	714 (595, 836)	0.82	780 (654, 916)	0.86	907 (753, 1067)	0.88	1111 (932, 1296)
3	5	0.74	461 (418, 504)	0.76	468 (421, 515)	0.75	485 (439, 531)	0.79	518 (472, 568)	0.82	565 (512, 617)	0.85	637 (575, 700)
	0.5	0.74	455 (408, 503)	0.74	460 (413, 508)	0.77	477 (429, 525)	0.78	509 (458, 562)	0.82	555 (500, 613)	0.86	627 (565, 691)
	0.3	0.75	442 (391, 498)	0.76	448 (394, 504)	0.77	466 (410, 524)	0.79	496 (439, 556)	0.81	541 (476, 606)	0.86	614 (544, 688)
	0.2	0.76	417 (343, 495)	0.75	424 (349, 502)	0.77	438 (358, 519)	0.79	466 (380, 555)	0.82	511 (421, 607)	0.86	573 (467, 683)
4	5	0.67	459 (415, 501)	0.68	466 (422, 511)	0.70	482 (431, 530)	0.72	513 (467, 563)	0.74	557 (505, 612)	0.76	636 (576, 694)
	0.5	0.66	453 (406, 502)	0.69	459 (412, 507)	0.70	476 (430, 524)	0.71	506 (452, 557)	0.74	555 (498, 612)	0.76	623 (560, 688)
	0.3	0.69	440 (386, 496)	0.69	446 (391, 503)	0.70	461 (402, 523)	0.71	491 (431, 555)	0.74	536 (469, 604)	0.77	607 (529, 686)
	0.2	0.70	417 (337, 504)	0.71	419 (337, 506)	0.72	435 (350, 523)	0.73	463 (377, 554)	0.75	507 (408, 607)	0.78	576 (468, 689)

Table A.7: Power and average sample size (first and third quartile in brackets) under Scenario 4 (binary outcome) computed using the Bayesian ‘two priors’ approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to 0.

		σ_d							
		0		0.01		0.02		0.03	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n
1	5	0.87	1843 (1532, 2191)	0.89	1975 (1646, 2345)	0.93	2443 (2028, 2903)	0.98	3637 (3045, 4306)
	0.5	0.87	1849 (1542, 2197)	0.89	1981 (1659, 2346)	0.93	2438 (2027, 2906)	0.98	3658 (3035, 4343)
	0.3	0.87	1856 (1541, 2206)	0.89	1978 (1644, 2358)	0.93	2447 (2036, 2904)	0.98	3636 (3025, 4317)
	0.1	0.86	1920 (1607, 2273)	0.90	2060 (1709, 2446)	0.93	2516 (2093, 2987)	0.98	3742 (3100, 4442)
2	5	0.83	1841 (1528, 2189)	0.84	1971 (1645, 2338)	0.89	2448 (2035, 2906)	0.95	3622 (3028, 4300)
	0.5	0.82	1846 (1540, 2190)	0.84	1973 (1650, 2342)	0.88	2444 (2046, 2898)	0.95	3644 (3019, 4326)
	0.3	0.82	1860 (1546, 2208)	0.84	1983 (1660, 2352)	0.88	2447 (2041, 2898)	0.94	3632 (3028, 4303)
	0.1	0.82	1912 (1598, 2270)	0.85	2052 (1700, 2436)	0.88	2512 (2095, 2977)	0.95	3720 (3074, 4408)
3	5	0.76	1182 (982, 1396)	0.74	1227 (1022, 1454)	0.79	1402 (1170, 1666)	0.86	1792 (1493, 2129)
	0.5	0.71	1167 (974, 1388)	0.75	1236 (1029, 1463)	0.79	1415 (1178, 1677)	0.86	1784 (1486, 2112)
	0.3	0.74	1194 (996, 1411)	0.74	1240 (1026, 1472)	0.80	1426 (1193, 1677)	0.86	1790 (1485, 2120)
	0.1	0.71	1250 (1045, 1476)	0.76	1303 (1080, 1548)	0.80	1479 (1236, 1743)	0.86	1869 (1560, 2206)
4	5	0.70	1190 (982, 1414)	0.72	1242 (1026, 1468)	0.77	1422 (1182, 1680)	0.81	1796 (1488, 2130)
	0.5	0.70	1183 (979, 1406)	0.72	1237 (1027, 1460)	0.75	1410 (1170, 1666)	0.82	1795 (1484, 2139)
	0.3	0.70	1192 (987, 1416)	0.72	1249 (1037, 1476)	0.76	1410 (1174, 1673)	0.82	1790 (1482, 2124)
	0.1	0.70	1248 (1040, 1476)	0.70	1297 (1080, 1531)	0.74	1498 (1248, 1770)	0.83	1854 (1529, 2210)

Table A.8: Power and average sample size (first and third quartile in brackets) under Scenario 4 (binary outcome) computed using the Bayesian 'two priors' approach for different values of σ_0 and σ_d. The analysis prior mean θ_0 is set to $\widehat{\mu}^p_{\bar{d}^1_k} - \widehat{\mu}^p_{\bar{d}^2_k}$ (estimated from the simulated pilot study).

		σ_d							
		0		0.01		0.02		0.03	
Setting	σ_0	Power	n	Power	n	Power	n	Power	n
1	5	0.87	1843 (1532, 2191)	0.89	1975 (1646, 2345)	0.93	2443 (2028, 2902)	0.98	3637 (3045, 4306)
	0.5	0.87	1840 (1533, 2187)	0.89	1971 (1651, 2334)	0.93	2425 (2016, 2888)	0.98	3641 (3023, 4329)
	0.3	0.88	1830 (1520, 2183)	0.89	1948 (1620, 2325)	0.93	2413 (2013, 2866)	0.98	3590 (2987, 4262)
	0.2	0.87	1807 (1502, 2145)	0.89	1940 (1600, 2315)	0.94	2380 (1952, 2829)	0.98	3572 (2946, 4260)
2	5	0.83	1840 (1528, 2189)	0.84	1971 (1644, 2338)	0.89	2448 (2035, 2906)	0.95	3621 (3028, 4299)
	0.5	0.83	1836 (1536, 2182)	0.84	1963 (1644, 2333)	0.88	2431 (2035, 2886)	0.94	3627 (3011, 4302)
	0.3	0.83	1833 (1525, 2183)	0.84	1953 (1637, 2322)	0.88	2413 (2016, 2863)	0.94	3584 (2988, 4249)
	0.2	0.83	1798 (1488, 2141)	0.85	1933 (1586, 2308)	0.89	2375 (1955, 2823)	0.95	3550 (2910, 4224)
3	5	0.74	1181 (980, 1391)	0.74	1241 (1024, 1472)	0.79	1409 (1178, 1657)	0.86	1784 (1483, 2107)
	0.5	0.73	1177 (981, 1394)	0.75	1237 (1024, 1468)	0.79	1413 (1169, 1676)	0.87	1781 (1475, 2119)
	0.3	0.74	1179 (969, 1413)	0.74	1219 (1011, 1453)	0.79	1383 (1143, 1653)	0.86	1762 (1469, 2090)
	0.2	0.73	1149 (937, 1362)	0.75	1197 (962, 1432)	0.80	1366 (1100, 1621)	0.86	1731 (1411, 2061)
4	5	0.71	1183 (980, 1412)	0.71	1234 (1023, 1467)	0.75	1414 (1180, 1666)	0.82	1780 (1478, 2103)
	0.5	0.71	1189 (986, 1407)	0.72	1232 (1016, 1462)	0.77	1404 (1163, 1671)	0.81	1771 (1466, 2100)
	0.3	0.70	1168 (970, 1384)	0.72	1212 (1021, 1435)	0.76	1390 (1151, 1652)	0.82	1761 (1474, 2091)
	0.2	0.71	1144 (918, 1369)	0.71	1197 (969, 1422)	0.76	1367 (1093, 1629)	0.81	1731 (1416, 2059)

Table A.9: Simulated type I error under the proposed methodology for different choices of prior parameters in the continuous final outcome setting.

			σ_d			
			0	0.2	0.5	0.8
$\theta_0 = 0$	σ_0	100	0.042	0.038	0.043	0.040
		3	0.036	0.035	0.039	0.040
		2	0.037	0.035	0.040	0.037
		1	0.017	0.022	0.021	0.030
$\theta_0 = \widehat{\theta}^p$	σ_0	100	0.046	0.040	0.043	0.041
		5	0.042	0.040	0.038	0.043
		4	0.045	0.034	0.040	0.038
		3	0.040	0.040	0.035	0.040

Table A.10: Simulated type I error under the proposed methodology for different choices of prior parameters in the binary final outcome setting.

			σ_d					
			0	0.01	0.02	0.03	0.04	0.05
$\theta_0 = 0$	σ_0	5	0.052	0.050	0.055	0.057	0.055	0.050
		0.5	0.060	0.048	0.045	0.053	0.052	0.058
		0.3	0.045	0.044	0.048	0.052	0.048	0.039
		0.1	0.035	0.034	0.040	0.040	0.038	0.042
$\theta_0 = \widehat{\theta}^p$	σ_0	5	0.053	0.054	0.047	0.046	0.051	0.049
		0.5	0.061	0.051	0.042	0.059	0.049	0.053
		0.3	0.054	0.051	0.052	0.054	0.049	0.042
		0.2	0.049	0.056	0.055	0.053	0.054	0.060

APPENDIX B

bayesSMARTsize online tutorial for Manuscript 1

`bayesSMARTsize` is a package which implements the functions for the Bayesian 'two priors' approach sample size estimation of a 2-stages sequential multiple assignment randomized trial (SMART) with continuous outcomes for the comparison of two strategies with different initial treatments. For a full description of this methodology, please refer to Turchetta et al. [2022]. In particular, the SMART design for which sample size formulae are provided is the standard scheme represented below, where responders to the stage-1 intervention are not re-randomized.

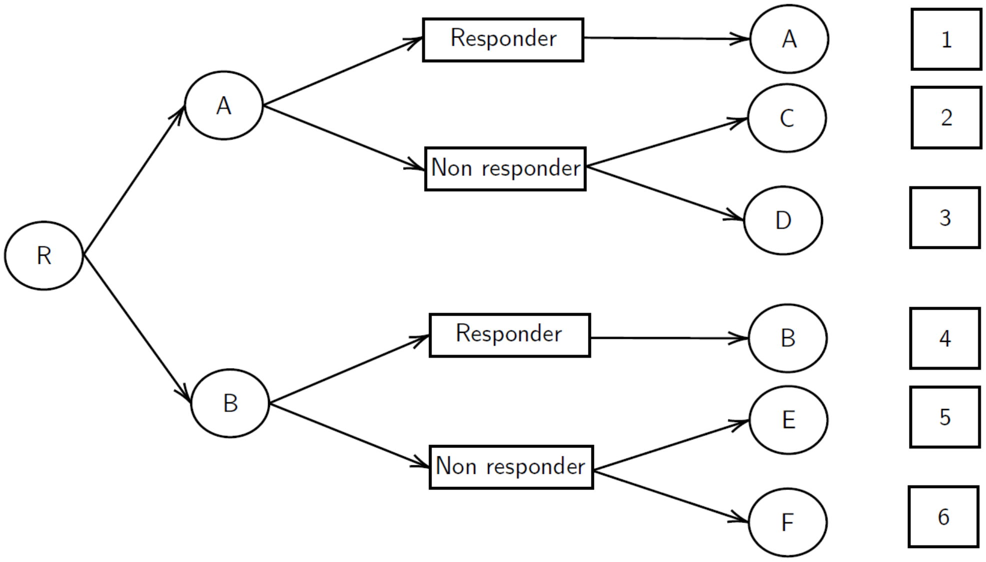

Figure B.1: SMART scheme.

Despite this methodology is built on the continuous outcome setting, it can still be used for binary outcomes, as the Normal approximation performs well.

,

```r
library(devtools)
install_github("aturchetta/bayesSMARTsize")
library(bayesSMARTsize)
set.seed(123)
```

B.1 Data generating functions

`SMART.continuous` and `SMART.binary` can be used to simulate data from a SMART with continuous and binary final outcomes respectively. Considering the labels depicted in the figure above, the response rates to the initial treatments $\mathtt{p}_{\mathtt{A}}$ and $\mathtt{p}_{\mathtt{B}}$ need to be specified in both cases, however, generating data in the binary final outcome setting is easier, as it only requires the specification of the response rates to the various treatment sequences. See the example below for a SMART with a sample size of 80 individuals.

,

```
sim_bin <- SMART.binary(n = 80, p_A = 0.3, p_AC = 0.4, p_AD = 0.3, p_AA = 1,
                        p_B = 0.3, p_BE = 0.2, p_BF = 0.2, p_BB = 1 )
```

In the continuous setting, the outcome is drawn from a Normal distribution with mean $\mathrm{E}[Y|A_1, R, A_2]$ and variance $\mathrm{Var}[Y|A_1, R, A_2]$, where A_j represent the treatment assigned at stage j and R the response to the stage-1 intervention (1 for responders and 0 for non-responders). Expressing the conditional mean of the outcome as

$$\begin{aligned}\mathrm{E}[Y|A_1, R, A_2] = \phi_1 + \phi_2 1_{\{A_1=a\}} + \phi_3(1-R) + \phi_4 1_{\{A_1=a\}}(1-R) \\ + \phi_5 1_{\{A_2=c\,\cup\, A_2=e\}}(1-R) + \phi_6 1_{\{A_1=a\,\cup\, A_2=c\}}(1-R),\end{aligned}$$

the algorithm to simulate requires the specification of 12 parameters: 6 parameters for the conditional mean ($\texttt{phi}_1, \ldots, \texttt{phi}_6$) and 6 standard deviation values ($\texttt{sd}_{\texttt{aa}}, \texttt{sd}_{\texttt{ac}}, \texttt{sd}_{\texttt{ad}}, \texttt{sd}_{\texttt{bb}}, \texttt{sd}_{\texttt{be}}, \texttt{sd}_{\texttt{bf}}$). See the example below.

,

```
sim_cont <- SMART.continuous(n = 80, p_A = 0.5, p_B = 0.5, phi_1 = 10, phi_2 = 5,
   phi_3 = -15,
                          phi_4 = -3, phi_5 = 10, phi_6 = -3, sd_aa = 2, sd_ac =
                             2, sd_ad = 2,
                          sd_bb = 2, sd_be = 3, sd_bf = 2)
```

Both functions return a data frame with 4 columns: the stage-1 intervention label, the sequence of treatments label, the final outcome, and the response indicator to the first treatment are returned:

```
head(sim_bin)
#>   tr1 tr2 outcome R
#> 1   A  AD       0 0
#> 2   B  BE       0 0
#> 3   B  BE       0 0
#> 4   A  AD       0 0
#> 5   A  AD       1 0
#> 6   A  AD       1 0

head(sim_cont)
#>   tr1 tr2   outcome R
#> 1   B  BF -1.089412 0
#> 2   B  BB  8.779077 1
#> 3   A  AA 12.365968 1
#> 4   B  BE  5.972913 0
#> 5   B  BF -6.004397 0
#> 6   A  AA 15.925476 1
```

B.2 Estimator function

SMART.est can be used to estimate the mean $\widehat{\mu}_{\overline{d}_k}$ and its estimator's variance $\tau^2_{\overline{d}_k}$ of a strategy $\overline{d}_k$. Let us take as an example the data set generated in the previous section in the binary outcome setting and let us consider the strategy 'administer A and, if there is no response, administer C'. Since responders are not randomized again, only the label of the stage-2 intervention is required, and in this case the label is 'AC':

```
SMART.est(tr1 = sim_bin$tr1, tr2 = sim_bin$tr2, outcome = sim_bin$outcome,
          R = sim_bin$R, id1 = "A", id2 = "AC")
#> $mu
#> [1] 0.8461538
#>
#> $tau
#> [1] 0.4946746
```

B.3 Sample size estimation

`SMART.ss` computes the required sample size to achieve the desired level of power. Let us consider the system of hypotheses

$$\begin{cases} \mathrm{H}_0 : \mu_{\overline{d}_k^1} - \mu_{\overline{d}_k^2} = 0 \\ \mathrm{H}_1 : \mu_{\overline{d}_k^1} - \mu_{\overline{d}_k^2} > 0. \end{cases}$$

The Bayesian methodology implemented in this function is based on the use of the posterior distribution of the variance components estimated from pilot data to marginalize the power function. Therefore, this function requires data from the pilot study through the set of parameters {`tr1, tr2, outcome, R, id1_study, id2_study, id1_ref, id2_ref`}. Alternatively, one can directly specify the posterior parameters of the inverse chi-squared distribution of $\tau^2_{\overline{d}_k^1} + \tau^2_{\overline{d}_k^2}$ through the parameters `vn` and `sigma_n`. Note that only one of the two choices must be specified in the function. Let us consider the data simulated in the binary outcome setting as the result of a pilot study and let us size the full-scale SMART for the comparison of $\overline{d}_k^1$: 'administer A and, if there is no response, switch to C' and $\overline{d}_k^2$: 'administer B and, if there is no response, switch to E'. We size the SMART in order to achieve a 90% power under the design prior distribution of the mean difference between strategies $\pi_d(\theta) = \mathcal{N}\left(\theta; 0.1, 0.02^2\right)$ and we set the Bayesian significance level to 0.95 ('epsilon=0.05'). To complete the Bayesian framework, we set the analysis prior $\pi_0(\theta) = \mathcal{N}\left(\theta; 0, 0.15^2\right)$. For further details on this methodology, please consult Turchetta et al. [2022]. Finally, a grid of sample size values which is used to search for the optimal sample size needs to be specified through `n_grid`. By default, the function outputs the estimated sample size and the corresponding power:

,

```
SMART.ss(n_grid = seq(100,1200), theta_0=0, sigma_0 = 0.15,
               theta_d=0.1, sigma_d = 0.01, power = 0.9, epsilon = 0.05,
               tr1 = sim_bin$tr1, tr2 = sim_bin$tr2, outcome = sim_bin$outcome, R
                   = sim_bin$R,
               id1_study = "A", id2_study = "AB", id1_ref = "B", id2_ref = "BE")
#>   Sample_size    Power
#> 1         964 0.9002337
```

Selecting the option save_grid=TRUE, an additional data frame which includes the power level estimated for each value of n_grid is returned:

,

```
grid <- SMART.ss(n_grid = seq(400,1000), theta_0=0, sigma_0 = 0.15,
               theta_d=0.1, sigma_d = 0.01, power = 0.9, epsilon = 0.05,
               tr1 = sim_bin$tr1, tr2 = sim_bin$tr2, outcome = sim_bin$outcome, R
                   = sim_bin$R,
               id1_study = "A", id2_study = "AB", id1_ref = "B", id2_ref = "BE",
                   save_grid=TRUE)

head(grid[[2]])
#>     n     Power
#> 1 400 0.5896011
#> 2 401 0.5906061
#> 3 402 0.5916087
#> 4 403 0.5926091
#> 5 404 0.5936072
#> 6 405 0.5946029
```

Additionally, it is possible to center the analysis prior π_0 at the mean difference between strategies estimated via pilot data through the option theta_0="pilot". In this case, incrementing the variance of the analysis prior to 0.5^2, the sample size is reduced to

,

```
SMART.ss(n_grid = seq(100,1200), theta_0="pilot", sigma_0 = 0.5,
               theta_d=0.1, sigma_d = 0.01, power = 0.9, epsilon = 0.05,
               tr1 = sim_bin$tr1, tr2 = sim_bin$tr2, outcome = sim_bin$outcome, R
                   = sim_bin$R,
               id1_study = "A", id2_study = "AB", id1_ref = "B", id2_ref = "BE")
#>   Sample_size    Power
#> 1         885 0.9000118
```

APPENDIX C

Appendix to Manuscript 3

C.1 Estimated parameters in the HIV vaccine trials

Table C.1: Estimates of α and plateau point t_p, polynomial degree and internal knot position (if any) of the best-fitting model for each interim time in the HIV vaccine trials.

Study	t_{int}	$\widehat{\alpha}$	$\widehat{t_p}$	Pol. deg.	Int. knot
HVTN 702	250	0.42	203	2	$\widehat{t_p}/4$
	400	0.58	394	2	None
	550	0.59	543	2	$\widehat{t_p}/2$
	700	0.61	675	2	$\widehat{t_p}/4$
HVTN 705	150	0.68	101	3	None
	300	1.00	110	2	None
	450	0.83	350	3	$\widehat{t_p}/4$

C.2 Forecastings in HVTN 702 without breaks removal

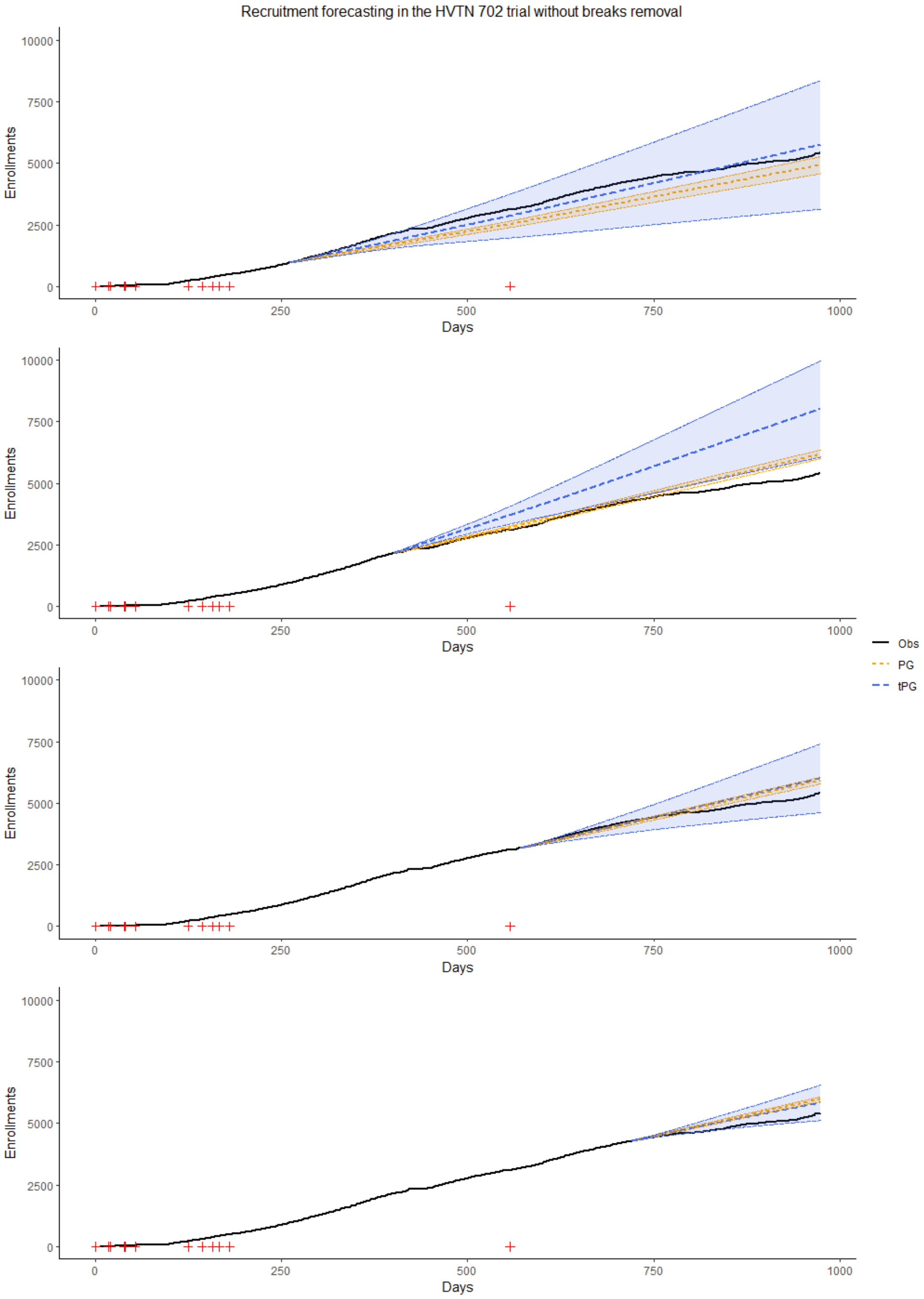

Figure C.1: Recruitment forecastings in the HVTN 702 study without the removal of the December holiday breaks under the tPG and standard PG models.

C.3 The PROBIT case study

The Promotion of Breastfeeding Intervention Trial (PROBIT) is a multicenter randomized cluster trial whose goal is to assess the effects of a breastfeeding promotion intervention designed to increase the duration and exclusivity of breastfeeding on maternal and child health outcomes. The trial was conducted in Belarus, where 17,046 mothers were recruited in 31 centers between June 1996 and December 1997. Figure C.2 displays the recruitment predictions under the tPG and standard PG models. Since the estimated plateau point is close to the initiation date of the centers (day 11 for both interim times), the tPG and PG models led to comparable point estimates and credible intervals.

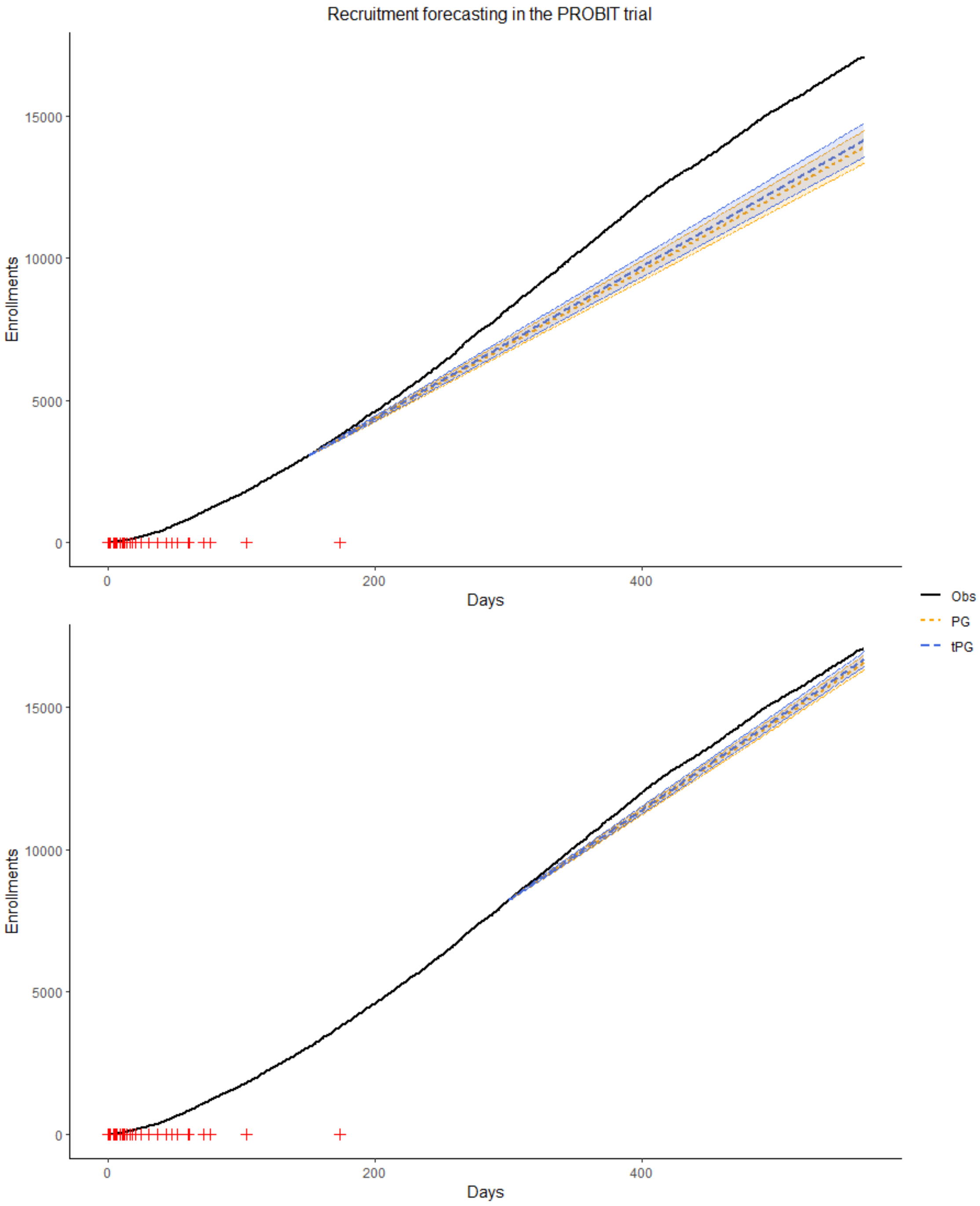

Figure C.2: Enrollment predictions in the PROBIT study under the time-dependent and standard PG models compared to the observed recruitments. The shaded areas indicate the respective 95% credible intervals and the + marks represent the centers' initiation dates.

References

C. J. Adcock. A Bayesian approach to calculating sample sizes. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37(4-5):433–439, 1988.

C. J. Adcock. Sample size determination: A review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):261–283, 1997.

M. Allen, H. Israel, K. Rybczyk, M. A. Pugliese, K. Loughran, L. Wagner, and S. Erb. Trial-related discrimination in HIV vaccine clinical trials. *AIDS Research and Human Retroviruses*, 17(8):667–674, 2001.

D. Almirall, S. N. Compton, M. Gunlicks-Stoessel, N. Duan, and S. A. Murphy. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in Medicine*, 31(17):1887–1902, 2012.

D. Almirall, I. Nahum-Shani, N. E. Sherwood, and S. A. Murphy. Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Translational Behavioral Medicine*, 4(3):260–274, 2014.

M. P. Andrasik, F. A. Sesay, A. Isaacs, L. Oseso, and M. Allen. Social impacts among participants in HIV Vaccine Trial Network (HVTN) preventive HIV vaccine trials. *Journal of Acquired Immune Deficiency Syndromes*, 84(5):488, 2020.

V. V. Anisimov. Using mixed Poisson models in patient recruitment in multicentre clinical trials. In *Proceedings of the World Congress on Engineering*, volume 2, pages 1046–1049, 2008.

V. V. Anisimov. Predictive modelling of recruitment and drug supply in multicenter clinical trials. *Proc. of Joint Statistical Meeting*, pages 1248–1259, 2009a.

V. V. Anisimov. Recruitment modeling and predicting in clinical trials. *Pharmaceutical Outsourcing*, 10(1):44–48, 2009b.

V. V. Anisimov. Statistical modeling of clinical trials (recruitment and randomization). *Communications in Statistics-Theory and Methods*, 40(19-20):3684–3699, 2011a.

V. V. Anisimov. Predictive event modelling in multicenter clinical trials with waiting time to response. *Pharmaceutical Statistics*, 10(6):517–522, 2011b.

V. V. Anisimov. Effects of unstratified and centre-stratified randomization in multi-centre clinical trials. *Pharmaceutical Statistics*, 10(1):50–59, 2011c.

V. V. Anisimov and V. V. Fedorov. Modeling of enrolment and estimation of parameters in multicentre trials. *GlaxoSmithKline Pharmaceuticals*, 66, 2005.

V. V. Anisimov and V. V. Fedorov. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine*, 26(27):4958–4975, 2007a.

V. V. Anisimov and V. V. Fedorov. Design of multicentre clinical trials with random enrolment. *Advances in Statistical Methods for the Health Sciences: Applications to Cancer and Aids Studies, Genome Sequence Analysis, and Survival Analysis*, pages 387–400, 2007b.

V. V. Anisimov, D. Downing, and V. V. Fedorov. Recruitment in multicentre trials: Prediction and adjustment. In *mODa 8-Advances in Model-Oriented Design and Analysis*, pages 1–8. Springer, 2007.

V. V. Anisimov, G. Mijoule, A. Turchetta, and N. Savy. Modelling and forecasting patient recruitment in clinical trials with patients' dropout. *arXiv preprint arXiv:2202.06779*, 2022.

W. J. Artman, I. Nahum-Shani, T. Wu, J. R. Mckay, and A. Ertefaie. Power analysis in a SMART design: Sample size estimation for determining the best embedded dynamic treatment regime. *Biostatistics*, 21(3):432–448, 2020.

W. J. Artman, B. A. Johnson, K. G. Lynch, J. R. McKay, and A. Ertefaie. Bayesian set of best dynamic treatment regimes: Construction and sample size calculation for SMARTs with binary outcomes. *Statistics in Medicine*, 2022.

S. F. Auyeung, Q. Long, E. B. Royster, S. Murthy, M. D. McNutt, D. Lawson, A. Miller, A. Manatunga, and D. L. Musselman. Sequential multiple-assignment randomized trial design of neurobehavioral treatment for patients with metastatic malignant melanoma undergoing high-dose interferon-alpha therapy. *Clinical Trials*, 6(5):480–490, 2009.

A. Bakhshi, S. Senn, and A. Phillips. Some issues in predicting patient recruitment in multi-centre clinical trials. *Statistics in Medicine*, 32(30):5458–5468, 2013.

K. D. Barnard, L. Dent, and A. Cook. A systematic review of models to predict recruitment to multicentre clinical trials. *BMC Medical Research Methodology*, 10:1–8, 2010.

M. L. Bell, A. L. Whitehead, and S. A. Julious. Guidance for using pilot studies to inform the design of intervention trials with continuous outcomes. *Clinical Epidemiology*, 10:153–157, 2018.

C. Bieganek, C. Aliferis, and S. Ma. Prediction of clinical trial enrollment rates. *Plos One*, 17(2):e0263193, 2022.

T. Bigirumurame, G. Uwimpuhwe, and J. Wason. Sequential multiple assignment randomized trial studies should report all key components: A systematic review. *Journal of Clinical Epidemiology*, 142:152–160, 2022.

V. Bogin. Lasagna's law: A dish best served early. *Contemporary Clinical Trials Communications*, page 100900, 2022.

R. H. Browne. On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17):1933–1940, 1995.

P. Brutti, F. De Santis, and S. Gubbiotti. Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine*, 27(13):2290–2306, 2008.

P. Brutti, F. De Santis, and S. Gubbiotti. Bayesian-frequentist sample size determination: A game of two priors. *Metron*, 72(2):133–151, 2014.

J. Cao, J. J. Lee, and S. Alber. Comparison of Bayesian sample size criteria: ACC, ALC, and WOC. *Journal of Statistical Planning and Inference*, 139(12):4111–4122, 2009.

R. E. Carter. Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials*, 25(5):429–436, 2004.

B. Chakraborty and E. E. M. Moodie. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

B. Chakraborty and S. A. Murphy. Dynamic treatment regimes. *Annual Review of Statistics and its Application*, 1:447–464, 2014.

Y. K. Cheung, B. Chakraborty, and K. W. Davidson. Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. *Biometrics*, 71(2):450–459, 2015.

ClinicalTrials.gov. A study to assess the efficacy of a heterologous prime/boost vaccine regimen of Ad26.Mos4.HIV and aluminum phosphate-adjuvanted clade C gp140 in preventing human immunodeficiency virus (HIV) -1 infection in women in Sub-Saharan Africa. URL https://clinicaltrials.gov/ct2/show/NCT03060629. accessed on April 17, 2023.

R. Dawson and P. W. Lavori. Sample size calculations for evaluating treatment policies in multi-stage designs. *Clinical Trials*, 7(6):643–652, 2010.

R. Dawson and P. W. Lavori. Efficient design and inference for multistage randomized trials of individualized treatment policies. *Biostatistics*, 13(1):142–152, 2012.

C. De Boor. *A practical guide to splines*, volume 27. Springer-Verlag, 1978.

F. De Santis. Sample size determination for robust Bayesian analysis. *Journal of the American Statistical Association*, 101(473):278–291, 2006.

Y. Deng, X. Zhang, and Q. Long. Bayesian modeling and prediction of accrual in multi-regional clinical trials. *Statistical Methods in Medical Research*, 26(2):752–765, 2017.

J. J. Dziak, D. Almirall, W. Dempsey, C. Stanger, and I. Nahum-Shani. SMART binary: Sample size calculation for comparing adaptive interventions in SMART studies with longitudinal binary outcomes. *arXiv preprint arXiv:2110.05535*, 2021.

A. Ertefaie, T. Wu, K. G. Lynch, and I. Nahum-Shani. Identifying a set that contains the best dynamic treatment regimes. *Biostatistics*, 17(1):135–148, 2016.

D. B. Fogel. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*, 11: 156–164, 2018.

B. Freedman. Equipoise and the ethics of clinical research. *New England Journal of Medicine*, 317(3):141–145, 1987.

S. S. Fu, A. J. Rothman, D. M. Vock, B. Lindgren, D. Almirall, A. Begnaud, A. Melzer, K. Schertz, S. Glaeser, P. Hammett, et al. Program for lung cancer screening and tobacco cessation: Study protocol of a sequential, multiple assignment, randomized trial. *Contemporary Clinical Trials*, 60:86–95, 2017.

B. J. Gajewski, S. D. Simon, and S. E. Carlson. Predicting accrual in clinical trials with Bayesian posterior predictive distributions. *Statistics in Medicine*, 27(13):2328–2340, 2008.

E. Gkioni, R. Rius, S. Dodd, and C. Gamble. A systematic review describes models for recruitment prediction at the design stage of a clinical trial. *Journal of Clinical Epidemiology*, 115:141–149, 2019.

E. Gkioni, S. Dodd, R. Rius, and C. Gamble. Statistical models to predict recruitment in clinical trials were rarely used by statisticians in UK and European networks. *Journal of Clinical Epidemiology*, 124:58–68, 2020.

G. E. Gray, L.-G. Bekker, F. Laher, M. Malahleha, M. Allen, Z. Moodie, N. Grunenberg, Y. Huang, D. Grove, B. Prigmore, et al. Vaccine efficacy of ALVAC-HIV and bivalent subtype C gp120–MF59 in adults. *New England Journal of Medicine*, 384(12):1089–1100, 2021.

G. Gresham, J. L. Meinert, A. G. Gresham, and C. L. Meinert. Assessment of trends in the design, accrual, and completion of trials registered in ClinicalTrials.gov by sponsor type, 2000-2019. *JAMA Network Open*, 3(8):e2014682–e2014682, 2020.

E. Hariton and J. J. Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13): 1716, 2018.

J. He, D. K. McClish, and R. T. Sabo. Evaluating misclassification effects on single sequential treatment in sequential multiple assignment randomized trial (SMART) designs. *Statistics in Biopharmaceutical Research*, 14(3):306–313, 2022.

P. Heindel, B. V. Dieffenbach, N. L. Freeman, K. L. McGinigle, and M. T. Menard. Central concepts for randomized controlled trials and other emerging trial designs. In *Seminars in Vascular Surgery*. Elsevier, 2022.

D. F. Heitjan, Z. Ge, and G.-s. Ying. Real-time prediction of clinical trial enrollment and event counts: A review. *Contemporary Clinical Trials*, 45:26–33, 2015.

L. Y. Inoue, D. A. Berry, and G. Parmigiani. Relationship between Bayesian and frequentist sample size determination. *The American Statistician*, 59(1):79–87, 2005.

R. M. Jacques, R. Ahmed, J. Harper, A. Ranjan, I. Saeed, R. M. Simpson, and S. J. Walters. Recruitment, consent and retention of participants in randomised controlled trials: A review of trials published in the National Institute for Health Research (NIHR) Journals Library (1997–2020). *BMJ Open*, 12(2):e059230, 2022.

Y. Jiang, S. Simon, M. S. Mayo, and B. J. Gajewski. Modeling and validating Bayesian accrual models on clinical data and simulations using adaptive priors. *Statistics in Medicine*, 34(4):613–629, 2015.

L. Joseph and P. Belisle. Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):209–226, 1997.

L. Joseph and D. B. Wolfson. Interval-based versus decision theoretic criteria for the choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46 (2):145–149, 1997.

C. Kasari, A. Kaiser, K. Goods, J. Nietfeld, P. Mathy, R. Landa, S. A. Murphy, and D. Almirall. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(6):635–646, 2014.

B. Kasenda, E. Von Elm, J. You, A. Blümle, Y. Tomonaga, R. Saccilotto, A. Amstutz, T. Bengough, J. J. Meerpohl, M. Stegert, et al. Prevalence, characteristics, and publication of discontinued randomized trials. *Journal of the American Medical Association*, 311(10): 1045–1052, 2014.

S. A. Kelleher, C. S. Dorfman, J. C. P. Vilardaga, C. Majestic, J. Winger, V. Gandhi, C. Nunez, A. Van Denburg, R. A. Shelby, S. D. Reed, et al. Optimizing delivery of a behavioral pain intervention in cancer patients using a sequential multiple assignment randomized trial SMART. *Contemporary Clinical Trials*, 57:51–57, 2017.

A. Kenny, A. Luedtke, O. Hyrien, Y. Fong, R. Burnham, J. Heptinstall, S. Sawant, S. Stanfield-Oakley, F. L. Omar, S. Khuzwayo, O. Dintwe, E. Borducchi, L. Pattacini, W. Willems, L. Lavreys, J. van Duijn, D. J. Stieh, F. Tomaka, M. G. Pau, G. E. Gray, S. Buchbinder, K. Mngadi, M. J. McElrath, L. Corey, D. H. Barouch, S. C. De Rosa, G. Ferrari, E. Andersen-Nissen, G. Tomaras, and P. B. Gilbert. Immune correlates analysis of the Imbokodo HIV-1 vaccine efficacy trial. 2022. URL https://programme.aids2022.org/Abstract/Abstract/?abstractid=12669. Conference abstract (AIDS 2022).

K. M. Kidwell. SMART designs in cancer research: Past, present, and future. *Clinical Trials*, 11(4):445–456, 2014.

K. M. Kidwell, N. J. Seewald, Q. Tran, C. Kasari, and D. Almirall. Design and analysis considerations for comparing dynamic treatment regimens with binary outcomes from sequential multiple assignment randomized trials. *Journal of Applied Statistics*, 45(9): 1628–1651, 2018.

H. Kim. A sample size calculator for SMART pilot studies. *SIAM Undergraduate Research Online*, 9:229–250, 06 2016. 10.1137/15S014058.

M. B. Klein, S. Saeed, H. Yang, J. Cohen, B. Conway, C. Cooper, P. Côté, J. Cox, J. Gill,

D. Haase, et al. Cohort profile: The Canadian HIV–hepatitis C co-infection cohort study. *International Journal of Epidemiology*, 39(5):1162–1169, 2010.

M. R. Kosorok and E. B. Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6:263–286, 2019.

M. R. Kosorok and E. E. M. Moodie. *Adaptive treatment strategies in practice: Planning trials and analyzing data for personalized medicine*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015. 10.1137/1.9781611974188. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974188.

M. S. Kramer, B. Chalmers, E. D. Hodnett, Z. Sevkovskaya, I. Dzikovich, S. Shapiro, J.-P. Collet, I. Vanilovich, I. Mezen, T. Ducruet, et al. Promotion of Breastfeeding Intervention Trial (PROBIT): A randomized trial in the Republic of Belarus. *Journal of the American Medical Association*, 285(4):413–420, 2001.

K. Kunzmann, M. J. Grayling, K. M. Lee, D. S. Robertson, K. Rufibach, and J. M. Wason. A review of Bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, 75(4):424–432, 2021.

D. Lai, L. A. Moyé, B. R. Davis, L. E. Brown, and F. M. Sacks. Brownian motion and long-term clinical trial recruitment. *Journal of Statistical Planning and Inference*, 93(1-2): 239–246, 2001.

S. D. Lambert, S. Grover, A. M. Laizner, J. McCusker, E. Belzile, E. E. M. Moodie, J. W. Kayser, I. Lowensteyn, M. Vallis, M. Walker, D. Da Costa, L. Pilote, C. Ibberson, J. Sabetti, and M. de Raad. Adaptive web-based stress management programs among adults with a cardiovascular disease: A pilot sequential multiple assignment randomized trial (SMART). *Patient Education and Counseling*, 2021.

Y. Lan, G. Tang, and D. F. Heitjan. Statistical modeling and prediction of clinical trial recruitment. *Statistics in Medicine*, 38(6):945–955, 2019.

L. Lasagna. Problems in publication of clinical trial methodology. *Clinical Pharmacology & Therapeutics*, 25(5part2):751–753, 1979.

P. W. Lavori and R. Dawson. Developing and comparing strategies: An annotated portfolio of designs. *Psychopharmacology Bulletin*, 34(1):13, 1998.

P. W. Lavori and R. Dawson. A design for testing clinical strategies: Biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38, 2000.

P. W. Lavori, R. Dawson, and T. B. Mueller. Causal estimation of time-varying treatment effects in observational studies: Application to depressive disorder. *Statistics in Medicine*, 13(11):1089–1100, 1994.

P. W. Lavori, R. Dawson, and A. Rush. Flexible treatment strategies in chronic disease: Clinical and research implications. *Biological Psychiatry*, 48(6):605–614, 2000.

Y. J. Lee. Interim recruitment goals in clinical trials. *Journal of Chronic Diseases*, 36(5): 379–389, 1983.

H. Lei, I. Nahum-Shani, K. Lynch, D. Oslin, and S. A. Murphy. A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology*, 8: 21–48, 2012.

Z. Li and S. A. Murphy. Sample size formulae for two-stage randomized trials with survival outcomes. *Biometrika*, 98(3):503–518, 2011.

D. V. Lindley. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):129–138, 1997.

Y. Liu, Y. Wang, and D. Zeng. Sequential multiple assignment randomization trials with enrichment design. *Biometrics*, 73(2):378–390, 2017.

G. Lorenzoni, E. Petracci, E. Scarpi, I. Baldi, D. Gregori, and O. Nanni. Use of Sequential Multiple Assignment Randomized Trials (SMARTs) in oncology: Systematic review of published studies. *British Journal of Cancer*, 128(7):1177–1188, 2023.

S. H. Lovibond and P. F. Lovibond. *Manual for the depression anxiety stress scales*. Psychology Foundation of Australia, 1996.

C. Manschot, E. Laber, and M. Davidian. Interim monitoring of sequential multiple assignment randomized trials using partial information. *arXiv preprint arXiv:2209.06306*, 2022.

L. Martin, M. Hutchens, C. Hawkins, and A. Radnov. How much do clinical trials cost. *Nature Reviews Drug Discovery*, 16(6):381–382, 2017.

J. McCusker, J. M. Jones, M. Li, R. Faria, M. J. Yaffe, S. D. Lambert, A. Ciampi, E. Belzile, and M. de Raad. CanDirect: Effectiveness of a telephone-supported depression self-care intervention for cancer survivors. *Journal of Clinical Oncology*, 39(10):1150–1161, 2021.

G. Mijoule, S. Savy, and N. Savy. Models for patients' recruitment in clinical trials and sensitivity analysis. *Statistics in Medicine*, 31(16):1655–1674, 2012.

N. Minois, V. Lauwers-Cances, S. Savy, M. Attal, S. Andrieu, V. V. Anisimov, and N. Savy. Using Poisson–gamma model to evaluate the duration of recruitment process when historical trials are available. *Statistics in Medicine*, 36(23):3605–3620, 2017a.

N. Minois, S. Savy, V. Lauwers-Cances, S. Andrieu, and N. Savy. How to deal with the Poisson-gamma model to forecast patients' recruitment in clinical trials when there are pauses in recruitment dynamic? *Contemporary Clinical Trials Communications*, 5:144–152, 2017b.

E. E. M. Moodie, B. Chakraborty, and M. S. Kramer. Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40(4): 629–645, 2012.

A. Morciano and M. Moerbeek. Optimal allocation to treatments in a sequential multiple assignment randomized trial. *Statistical Methods in Medical Research*, 30(11):2471–2484, 2021.

M. Moussa. Planning a clinical trial with allowance for cost and patient recruitment rate. *Computer Programs in Biomedicine*, 18(3):173–179, 1984.

K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, University of British Columbia, 2007.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

S. A. Murphy. A generalization error for Q-Learning. *Journal of Machine Learning Research*, 6:1073–1097, 08 2005a.

S. A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481, 2005b.

S. A. Murphy, M. J. van der Laan, J. M. Robins, and C. P. P. R. Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456): 1410–1423, 2001.

S. A. Murphy, K. G. Lynch, D. Oslin, J. R. McKay, and T. TenHave. Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence*, 88:S24–S30, 2007.

T. A. Murray, Y. Yuan, and P. F. Thall. A Bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association*, 113(523): 1255–1267, 2018.

I. Nahum-Shani, A. Ertefaie, X. Lu, K. G. Lynch, J. R. McKay, D. W. Oslin, and D. Almirall. A SMART data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction*, 112(5):901–909, 2017.

T. NeCamp, A. Kilbourne, and D. Almirall. Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: Regression estimation and sample size considerations. *Statistical Methods in Medical Research*, 26(4):1572–1589, 2017.

A. I. Oetting, J. Levy, R. Weiss, and S. A. Murphy. Statistical methodology for a SMART design in the development of adaptive treatment strategies. *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, 8:179–205, 2011.

S. B. Ogbagaber, J. Karp, and A. S. Wahed. Design of sequentially randomized trials for testing adaptive treatment strategies. *Statistics in Medicine*, 35(6):840–858, 2016.

W. E. Pelham Jr, G. A. Fabiano, J. G. Waxmonsky, A. R. Greiner, E. M. Gnagy, W. E. Pelham III, S. Coxe, J. Verley, I. Bhatia, K. Hart, et al. Treatment sequencing for childhood ADHD: A multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology*, 45(4):396–415, 2016.

H. Pezeshk. Bayesian techniques for sample size determination in clinical trials: A short review. *Statistical Methods in Medical Research*, 12(6):489–504, 2003.

M. J. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26, 2009.

J. M. Robins. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer, 1997.

J. M. Robins. Optimal structural nested models for optimal sequential decisions. In *Pro-*

ceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data, pages 189–326. Springer, 2004.

E. J. Rose, E. B. Laber, M. Davidian, A. A. Tsiatis, Y.-Q. Zhao, and M. R. Kosorok. Sample size calculations for SMARTs. *arXiv preprint arXiv:1906.06646*, 2019.

S. K. Sahu and T. M. Smith. A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169 (2):235–253, 2006.

V. Sambucini. Bayesian vs frequentist power functions to determine the optimal sample size: Testing one sample binomial proportion using exact methods. In J. P. Tejedor, editor, *Bayesian Inference*, chapter 5. IntechOpen, Rijeka, 2017. 10.5772/intechopen.70168. URL https://doi.org/10.5772/intechopen.70168.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

A. I. Scott, J. A. Levy, and S. A. Murphy. Evaluation of sample size formulae for developing adaptive treatment strategies using a SMART design. Technical Report 07-81, University Park, PA: The Pennsylvania State University, The Methodology Center., 04 2007.

N. J. Seewald, K. M. Kidwell, I. Nahum-Shani, T. Wu, J. R. McKay, and D. Almirall. Sample size considerations for comparing dynamic treatment regimens in a sequential multiple-assignment randomized trial with a continuous longitudinal outcome. *Statistical Methods in Medical Research*, 29(7):1891–1912, 2020.

S. Senn. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine*, 17(15-16):1753–1765, 1998.

S. M. Shortreed, E. Laber, T. Scott Stroup, J. Pineau, and S. A. Murphy. A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine*, 33(24):4202–4214, 2014.

A. Sikorskii, G. Wyatt, R. Lehto, D. Victorson, T. Badger, and T. Pace. Using SMART design to improve symptom management among cancer patients: A study protocol. *Research in Nursing & Health*, 40(6):501–511, 2017.

D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons, 2004.

C. Stanger, E. A. Scherer, H. T. Vo, S. F. Babbin, A. A. Knapp, J. R. McKay, and A. J. Budney. Working memory training and high magnitude incentives for youth cannabis use: A SMART pilot trial. *Psychology of Addictive Behaviors*, 34(1):31, 2020.

T. S. Stroup, J. P. McEvoy, M. S. Swartz, M. J. Byerly, I. D. Glick, J. M. Canive, M. F. McGee, G. M. Simpson, M. C. Stevens, and J. A. Lieberman. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin*, 29(1):15, 2003.

G. Tang, Y. Kong, C.-C. H. Chang, L. Kong, and J. P. Costantino. Prediction of accrual closure date in multi-center clinical trials with discrete-time Poisson process models. *Pharmaceutical Statistics*, 11(5):351–356, 2012.

R. K. Tsutakawa. Design of experiment for bioassay. *Journal of the American Statistical Association*, 67(339):584–590, 1972.

A. Turchetta, E. E. M. Moodie, D. A. Stephens, and S. D. Lambert. Bayesian sample size calculations for comparing two strategies in SMART studies. *Biometrics*, 2022.

A. Turchetta, N. Savy, D. A. Stephens, E. E. Moodie, and M. B. Klein. A time-dependent

Poisson-Gamma model for recruitment forecasting in multicenter studies. *Statistics in Medicine*, 2023.

S. Urbas, C. Sherlock, and P. Metcalfe. Interim recruitment prediction for multi-center clinical trials. *Biostatistics*, 23(2):485–506, 2022.

J. C. van der Wouden, A. H. Blankenstein, M. J. Huibers, D. A. van der Windt, W. A. Stalman, and A. P. Verhagen. Survey among 78 studies showed that Lasagna's law holds in dutch primary care research. *Journal of Clinical Epidemiology*, 60(8):819–824, 2007.

A. J. Vickers. Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology*, 56(8):717–720, 2003.

M. P. Wallace and E. E. M. Moodie. Personalizing medicine: A review of adaptive treatment strategies. *Pharmacoepidemiology and Drug Safety*, 23(6):580–585, 2014.

M. P. Wallace and E. E. M. Moodie. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71(3):636–644, 2015.

M. P. Wallace, E. E. M. Moodie, and D. A. Stephens. SMART thinking: A review of recent developments in sequential multiple assignment randomized trials. *Current Epidemiology Reports*, 3(3):225–232, 2016.

S. J. Walters, I. B. dos Anjos Henriques-Cadby, O. Bortolami, L. Flight, D. Hind, R. M. Jacques, C. Knox, B. Nadin, J. Rothwell, M. Surtees, et al. Recruitment and retention of participants in randomised controlled trials: A review of trials funded and published by the United Kingdom Health Technology Assessment Programme. *BMJ Open*, 7(3): e015276, 2017.

F. Wang, A. E. Gelfand, et al. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2):193–208, 2002.

J. Wang, L. Wu, and A. S. Wahed. Adaptive randomization in a two-stage sequential multiple assignment randomized trial. *Biostatistics*, 23(4):1182–1199, 2022.

L. Wang, A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, 107(498):493–508, 2012.

X. Wang and B. Chakraborty. The sequential multiple assignment randomized trial for controlling infectious diseases: A review of recent developments. *American Journal of Public Health*, 113(1):49–59, 2023.

L. Wu, J. Wang, and A. S. Wahed. Interim monitoring in sequential multiple assignment randomized trials. *Biometrics*, 79(1):368–380, 2023.

X. Yan, P. Ghosh, and B. Chakraborty. Sample size calculation based on precision for pilot sequential multiple assignment randomized trial (SMART). *Biometrical Journal*, 63(2): 247–271, 2021.

J. Ypma, H. W. Borchers, D. Eddelbuettel, and M. J. Ypma. Package 'nloptr', 2018.

Q. Zhang and D. Lai. Fractional Brownian motion and long term clinical trial recruitment. *Journal of Statistical Planning and Inference*, 141(5):1783–1788, 2011.

X. Zhang and B. Huang. A simple and robust model for enrollment projection in clinical trials. *Contemporary Clinical Trials*, 123:106999, 2022.

X. Zhang and Q. Long. Stochastic modeling and prediction for accrual in clinical trials. *Statistics in Medicine*, 29(6):649–658, 2010.

X. Zhang and Q. Long. Joint monitoring and prediction of accrual and event times in clinical trials. *Biometrical Journal*, 54(6):735–749, 2012.