Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics and Computer Science

Cycle: 36th Cycle

Disciplinary Field (code): SECS-S/01

Methodologies for complex health economic modeling

Advisor: Marco Bonetti

Co-Advisor: Anna Heath

PhD Thesis by:

Luca Benetti

ID number: 3147447

Year 2026

Contents

Abstract

Making effective decisions in health and medicine is crucial, as each decision affects the well-being of others. Making cost-effective decisions in health and medicine is even more critical as economic resources are not infinite, and maximizing their utility is necessary. Health economic modeling refers to the process of evaluating the costs and effects of healthcare interventions. Decision analytic models are mathematical tools to account for and model the uncertainty around each decision, and to determine the best decision after collecting different sources of evidence. There are different open challenges and interesting problems related to statistical techniques to analyze complex and realistic evidence sources, as well as the uncertainty surrounding decision-making. In this thesis, we focus on the development of various Bayesian methodologies in complex and realistic frameworks within the context of health economic modeling.

In the first chapter, we provide an extensive introduction to the context of health economic modeling and cost-effective analysis, and outline the open challenges we will address throughout the thesis.

In Chapter 2, we define *Inverse target trial emulation*, a novel Bayesian methodology to generate realistic observational data. The central idea of this methodology is to start with an initial (preliminary) Randomized Clinical Trial (RCT) and use the initial information to generate different types of observational data in various contexts and under different assumptions. *Target trial emulation* (TTE) is a methodology that links an observational dataset to a targeted RCT, emulating experimental data by solving all the different forms

of bias in observational data. In this context, we reverse this process, aiming to simulate (not only emulate) observational data from an initial trial. Here, emulate means to estimate causal effects using real-world data in a way that mimics a randomized trial without explicitly simulating RCT data. This methodology proves to be very useful for performing different forms of sensitivity analysis, research prioritization, and testing of the robustness of the methods typically employed to analyze observational data.

In the third and fourth chapters, we focus our attention on VoI (Value of Information) analysis in complex and realistic scenarios. Given an underlying health-economic decision model, VoI analysis is a technique to quantify the expected benefit that may result from reducing uncertainty in the economic model. In particular, the Expected Value of Sample Information (EVSI) is a measure that estimates the expected benefit of performing additional data to reduce the uncertainty in the parameters of the underlying health economic model. Until now, the EVSI methodology has been applied only to fairly simple data collection exercises. In most cases, it has been used to understand the value of randomized clinical trial (RCT) data, meaning that it measured the value of collecting additional RCTs to reduce uncertainty. In these chapters, relying also on the results from Chapter 1, we design and apply a novel methodology to use EVSI when we plan to collect more complex and realistic data, i.e., data affected by missingness and confounding.

In the final chapter, we develop a Bayesian version of the *Subpopulation treatment effect pattern plot* (STEPP) methodology and apply it to real scenarios. STEPP is a methodology that enables researchers to properly analyze the heterogeneity of treatment effects (HTE) in experimental studies, providing the necessary information to customize treatment for individuals to maximize benefits. While traditional statistical methods for subgroup analysis divide the population into disjoint subgroups, STEPP takes a different approach by constructing overlapping subpopulations along the continuum of a continuous covariate of interest (e.g., a biomarker), thus improving the precision of the estimated treatment effects within the subgroups. In that chapter, we introduce a Bayesian version

of the STEPP method (B-STEPP) and demonstrate that a Bayesian approach enables flexible modeling of the dependence among the relevant parameters, providing good control over the joint distribution of the parameters and their associated uncertainty.

As we will prove, this thesis contributes to the field of statistical methods for health economics both by defining a suitable candidate for a unified approach to test causal inference methods and simulate observational data (ITTE), and by extending the use of powerful statistical techniques to more realistic contexts (realistic EVSI) and with more flexible approaches (B-STEPP).

Chapter 1

Introduction and motivating problems

1.1 Health economic modeling

Making decisions is always difficult, but it is necessary. It is particularly crucial in health economics, as decisions in this context significantly impact the well-being of others. In this context, health technology assessment (HTA) is a systematic, multidisciplinary process that evaluates the impact of a novel health technology, considering various factors such as its medical efficacy, cost-effectiveness, and safety.

A source of complexity in making decisions in health economics is that, in the real world, resources are not infinite. A decision maker must therefore evaluate different decision options to choose the one that optimizes the economic assets in relation to maximizing the utility of the defined outcomes (Hunink et al. (2014), Fenwick et al. (2000)). This challenge is addressed by *economic evaluation*, which relies on *expected utility theory* (Savage (1972), Raiffa (1968), Parmigiani and Inoue (2009)) to define a formal setting that allows for choosing among different decisions with associated costs and utilities. In health and medicine, the most common technique for economic evaluation is *cost-effectiveness*

analysis (CEA).

In a CEA, effectiveness can be measured in various ways, such as life expectancy, recurrence-free survival time, and health-related quality of life (HRQoL) (Neumann et al. (2016), Hunink et al. (2014)), among others. In addition, effectiveness measures can be defined as either health-specific or generic (Heath et al. (2024)). Generic measures have the advantage of ensuring comparability between decisions belonging to different contexts. Suppose, for example, that a decision maker has to choose which of two novel treatments to fund. If the treatments belong to two different contexts and have different clinical outcomes, then the only way to make a decision is to ensure comparability by defining a generic measure for effectiveness. The prevalent generic measure of effectiveness in health-related decision problems is the quality-adjusted life year (QALY). QALY is a way of measuring utility defined as the integral of quality of life (HRQoL) over the lifespan. In the computation of QALY, each health state is linked to individuals' perceived health-related quality of life; the latter are usually estimated as a number between 0 (death) and 1 (perfect health), in various ways (Heath et al. (2024)).

With respect to costs, the usual types considered in a CEA analysis are the costs associated with the consumption of the novel health technology and those related to the consequences of the decision (Neumann et al. (2016), Heath et al. (2024)). Moreover, the considered costs can be related solely to the healthcare sector or also to external factors and evaluated under different perspectives (societal and healthcare sector perspectives) (Baracskay (1998), Walker et al. (2019)).

In non-trivial situations, a given decision has better effects but also higher costs compared to another option. In this case, to determine the best decision, we must rely on the concept of *willingness-to-pay threshold* (WTP). WTP accounts for how much society is willing to pay for additional health gains (Claxton et al. (2015), Thokala et al. (2018)).

Given all these ingredients, the value of a decision in a health modeling context is commonly measured as the *net benefit* (Stinnett and Mullahy (1998)), that is:

$$\text{NB}_d = k \times e_d - c_d \tag{1.1}$$

With e_d and c_d, respectively, the effects and costs of decision d; while k is the cost that the decision maker is willing to pay for a unit of health effect. In this framework, a decision maker chooses the action d that maximizes NB_d.

1.1.1 Decision analytic model

In a health economic model, e_d and c_d are not deterministic, but exhibit different degrees of uncertainty. They depend mostly on the health status and on the clinical events individuals may experience after decision d has been taken. Modeling the various sources of uncertainty that characterize e_d and c_d is the main objective of a *decision analytic model*.

Decision analytic models are mathematical tools designed to model the consequences of different decisions on individuals, by defining the various states and the different clinical outcomes individuals may experience, which must be evaluated to inform a Cost-Effectiveness Analysis. There are different types of health decision analytic models; the most common ones are decision trees and state transition models, each of those are better suited to model different scenarios. Decision trees are generally well-suited for modeling simple health decision problems with a short-term horizon. In contrast, state transition models are better at evaluating more complex health decisions by modeling the individuals' state trajectories during follow-up and the effects of potential decisions. At the end of this section, we will introduce a specific health decision model that combines both a Markov state transition model and a decision tree (Heath et al. (2024)).

In general, decision analytic models are usually characterized by a vector of random parameters $\boldsymbol{\theta}$ with probability distributions $p(\boldsymbol{\theta})$, and D different possible decisions, $d = 1, \ldots, D$ a decision maker can choose from. The parameters represent various relevant

population-level clinical and non-clinical quantities, such as recurrence-free survival time under different therapies, the relative efficacy of treatments, disease prevalence, and other relevant measures. The probability distribution p characterizes the uncertainty around the population level parameters $\boldsymbol{\theta}$ (Briggs et al. (2012)), namely the population-level uncertainty.

The process of decision-making must change when we incorporate uncertainty into our framework. In a Bayesian decision framework, the probability distribution $p(\boldsymbol{\theta})$ is defined as the *prior distribution* on parameters $\boldsymbol{\theta}$, that is, the distribution that incorporates our prior knowledge on the uncertainty of all the relevant parameters. In this setting, since we do not know the actual net benefit for each decision, to choose the best decision, instead of maximizing equation (1.1) with respect to d, we would maximize the *expected net benefit*:

$$\mathbb{E}(\mathrm{NB}_d(\boldsymbol{\theta})) = k \times \mathbb{E}(e_d(\boldsymbol{\theta})) - \mathbb{E}(c_d(\boldsymbol{\theta})) \tag{1.2}$$

Concretely, the above maximization is performed by first sampling N values from $p(\boldsymbol{\theta})$ and then computing $\mathbb{E}(\mathrm{NB}_d(\boldsymbol{\theta}))$ via Monte Carlo (MC) simulation (Metropolis and Ulam (1949)). This process is known as *probabilistic sensitivity analysis* (PSA), or, as better named in (Heath et al. (2024)), *probabilistic analysis* (PA).

Modeling uncertainty in a Bayesian framework is a very useful approach for various reasons. First, it enables us to incorporate all prior knowledge into our model by defining the prior distribution p. Second, it allows us to incorporate any additional piece of information (data) we may collect in the future, allowing us to update the prior distribution naturally. This is achieved by computing the posterior distribution of relevant parameters given the additional data we collect, $\boldsymbol{\theta}|Y$, relying on Bayes' theorem (Van de Schoot et al. (2014)). Incorporating additional knowledge and obtaining the posterior distribution of $\boldsymbol{\theta}$ can also be very useful to investigate the convenience of collecting new data to reduce uncertainty and make more informed decisions; we will deepen this scenario when we

introduce *Value of Information* (VoI).

We now introduce the decision-analytic model from (Heath et al. (2024)), which will play a crucial role throughout this work, as we will apply the novel methodology of Chapters 3 and 4 to this model.

1.1.2 A Novel Chemotherapy Treatment

In this example, we suppose that a decision maker faces $D = 2$ different choices related to different chemotherapy treatments. Therapies do not differ in terms of efficacy, but they differ only in terms of costs and side effects profiles; obviously, to make it not a straightforward decision, we need one therapy to show higher costs for fewer side effects. Therefore, the research question is: Would the lower costs associated with fewer side effects justify the additional costs of the proposed therapy?

In this setting, the measure of health value is the Quality Adjusted Life Year (QALY), and all costs are measured in British pounds. The model we propose combines two common structures: a decision tree model and a Markov state transition model. The decision tree is used to represent the expected number of individuals who experience side effects, while the Markov model represents the possible trajectories of the health state and their related costs that individuals with side effects might incur. Figure 1.1 represents the four-state Markov model.

Initially, all individuals experiencing side effects are advised to manage them at home. After a certain time, if the side effects worsen, individuals may need to be treated in the hospital. We assume that only individuals who require hospitalization may die because of the side effects. For both individuals treated at home and in hospitals, there is a chance of recovery, and once recovered, we assume that an individual cannot experience side effects again.

The Markov model assumes weekly cycles and a one-year time horizon. At the end of the year, it is assumed that no individuals experience side effects anymore; that is, they

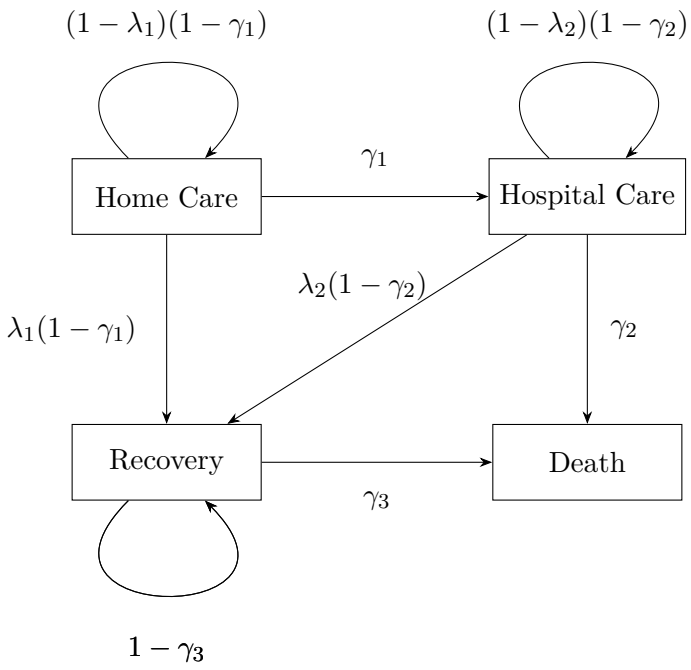


Figure 1.1: The four-state Markov transition model to compute costs and health effects for patients who experience side effects.

either recovered or died. Patients who remain alive after one year continue to be at risk of dying from cancer or other general causes.

Other relevant assumptions are: first, after individuals recover from side effects, they incur no further costs, and they have the same quality of life as individuals who do not experience side effects. Second, we assume that home care is less expensive and has better HRQOL compared to hospital treatment. Last, we assume that individuals who die from either side effects or other general causes have zero costs and zero associated health utility.

1.1.3 Model parameters and probabilistic analysis

The parameters associated with the decision tree are s_d for $d = 1, 2$ and model the proportion of people who experience side effects for standard and novel treatments. Since, in common scenarios, the baseline proportions of the standard treatment can be estimated

from the literature or a previous study, we focus our attention on the odds ratio

$$\rho = \frac{logit(s_1)}{logit(s_0)}.$$

Regarding the individuals who do not experience side effects, we assume a random quantity q to measure the quality of life throughout the 1-year horizon related to the Markov model. Moreover, we assume that the standard treatment has a cost of £120 while the novel one has a cost of £1975.

Since the definitions of transition probabilities and, in general, parameters related to the Markov state transition model are quite understandable from the Markov scheme above, we will focus directly on the forms of the probability distributions of the different relevant model parameters. These are necessary to perform probabilistic analysis in our framework.

Probability distributions of relevant parameters can be informed by either the literature or data sources. In the first case, we would utilize all the information from the relevant literature to select a suitable prior distribution for the model parameters. In the second case, a proper prior must be chosen based on the current level of information, and we would utilize data sources to update our prior and perform probabilistic analysis by sampling from the posterior distribution.

First, we list in Table 1.1 all the parameters informed by relevant literature, with related choices of prior distributions.

Parameter	Definition	Distribution	Estimate	Standard Error
$\log(\rho)$	Log odds ratio of side effects	Normal	log(0.54)	0.3
r	Rate of death for individuals who have recovered or not experienced side effects	Gamma	0.0475	0.0316
λ_1	Probability of recovery in a given week for someone treated at home who does not transition to hospital care	Beta	0.21	0.03
λ_2	Probability of recovery in a given week for someone treated in hospital who does not die	Beta	0.03	0.0065
q	Quality of life for recovered patients	Beta	0.98	0.0283
q_{HC}	Quality of life for home care patients	Beta	0.7	0.141
q_H	Quality of life for hospitalised patients	Beta	0.03	0.173
c_{death}	One-off cost of death	Log-normal	1710	27.57
c_{HC}	Yearly cost of treatment at home	Log-normal	830	12.25
c_H	Yearly cost of treatment in hospital	Log-normal	2400	43.36

Table 1.1: Distribution for parameters of Markov state transition model informed by the literature

Then, we move on and list in Table 1.2 all the relevant parameters that are informed by primary data sources. In the third column, we report the choice of the prior distribution

that will be updated.

Parameter	Definition	Prior Distribution	Prior Parameters	Data
s_0	Probability of side effects under standard care	Beta	(1, 1)	N: 111; Side Effects: 52
Γ_1	1-year probability of hospitalisation, given the patient had side effects	Beta	(1, 1)	Side Effects: 52; Hospitalisations: 43
Γ_2	1-year probability of death	Beta	(1, 1)	Hospitalisations: 43; Death: 8

Table 1.2: Prior distributions and data for the Chemotherapy model parameters

For a full specification on how posterior distributions are computed in this context and to understand how to transform year parameters such as Γ_1 into weekly transition probabilities such as γ_1, see (Heath et al. (2024)).

1.2 Evidence for health economic modeling

An important aspect that we have not explored up to now is related to the definition of the different sources of evidence that can be exploited to inform the relevant population-level parameters in decision analytic models.

In health economics, data collected from a given study can be divided into two main categories: experimental and observational data (O'Brien et al. (1994)). While experi-

mental data are collected from a deliberate experiment with a prespecified design, observational data are collected from routine activities in the real world. These two types differ significantly in structure and statistical characteristics, showing both advantages and disadvantages in their application to decision-making in HTA.

In general, experimental data have higher internal validity, meaning that randomization helps guarantee that differences in outcomes can be attributed to the intervention rather than various forms of bias (Cook et al. (2002)). Relatedly, experimental studies, if properly designed and conducted, are the most effective way to test causal assumptions. To this aim, the most popular type of experimental study is *Randomized Controlled Trials* (RCTs) (Hariton and Locascio (2018)), in which individuals are randomly assigned to treatment or control groups. Observational data, on the other hand, have greater external validity, meaning that they are, in principle, more applicable to clinical practice (Carlson and Morrison (2009)); however, they require additional assumptions and adjustments to answer causality questions.

Both observational and experimental studies play a crucial role in health technology assessment. Our aim is to make novel methodological contributions to both areas, introducing innovative methods to enhance the analysis of observational and experimental data and facilitate better decision-making in HTA.

1.2.1 Observational data

Even if experimental studies, and in particular RCTs, are still considered to be the gold standard for decision-making in HTA, they have several drawbacks. First, they do not ensure external validity because of the nature of experimental data. Second, experiments usually require very high associated costs to be conducted. Finally, in experiments, a problem of feasibility can arise for many different reasons. A simple example is that of rare diseases, where enrolling a large number of patients can be very challenging.

Observational data, on the other hand, have their own potential drawbacks as well.

Since they are collected primarily through routine healthcare delivery, they intrinsically contain different types of bias that can be related to one or more of the following broad categories (Grimes and Schulz (2002)):

- *Selection bias* refers to the bias that results when the selection of the study population fails to recreate proper randomization and to represent the target population.
- *Confounding* refers to the bias that arises when the association between exposure and outcome is distorted by the presence of a third factor, associated with both.
- *Information bias* refers to the bias related to systematic errors in the collection, registration, or classification of different variables.

These broader families of bias can manifest in different ways within observational data. Throughout this work, we will primarily focus on two common situations.

1.2.1.1 Imbalance in populations

The first bias that we refer to is the one related to **imbalance in populations**. As we already mentioned, randomization in RCTs enables us to mitigate selection bias and confounding by randomly allocating treatment. For this reason, the treated and control populations are asymptotically balanced in terms of measured and unmeasured personal characteristics (covariates), as they share the same probability distribution. Conversely, since in observational data allocation is not randomized, there is a risk of providing one treatment to some individuals and another treatment to quite different individuals. If the individual characteristics that differ in the two populations have a non-null correlation with the outcome variable, this may lead to a wrong estimation of treatment efficacy (Cochran and Chambers (1965)). In this case, we say that the imbalance in populations originated from confounding and/or selection bias, leading to a biased estimate of the efficacy of treatment. There are various ways to measure population imbalance, or overlap

in populations (standardized mean difference, propensity scores,...) (Imbens and Rubin (2015)). Throughout this work, we will measure it through the population Mahalanobis distance (McLachlan (1999)) as:

$$\Delta = \sqrt{(\boldsymbol{\mu}_t - \boldsymbol{\mu}_c)^T \left(\frac{\boldsymbol{S}_c + \boldsymbol{S}_t}{2}\right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{\mu}_c)}$$

The Mahalanobis distance between the mean values μ_t, μ_c of the treated and control populations with respect to the inner product $\left((\boldsymbol{S}_c + \boldsymbol{S}_t)/2\right)^{-1}$, where $\boldsymbol{S}_t, \boldsymbol{S}_c$ are the sample covariates matrix, respectively, of the treated and control individuals; and μ_t and μ_c represent the mean of the covariates distributions for the treatment and control group respectively.

As we will explain in greater detail in the following chapters, throughout the work we will also apply the pointwise version of Mahalanobis distance that measures the distance between two individuals i and j defined as:

$$d^2 = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the covariate profiles of i and j and $\boldsymbol{S}^{-1}$ is the inverse of the sample covariance matrix.

Recall that the Mahalanobis distance can be seen as a multivariate extension of the Euclidean distance that accounts for both the scale and the correlation structure of the covariates. In particular, it measures how many multivariate standard deviations apart two points (or two group means) are. Intuitively, in Mahalanobis distance, differences along highly variable directions are lower, while differences along tightly clustered or correlated dimensions are higher. In practice, it measures the Eucledian distances after rotating and re-scaling the axes through the inverse square root variance-covariance matrix. This makes the Mahalanobis distance a natural choice to measure covariate imbalance.

1.2.1.2 Immortal time bias

The second bias we refer to is linked to the definition of time zero. In non-experimental studies, it is often the case that after an individual enters the study, they must wait a suitable amount of time before starting therapy and being then labeled as treated (Tamm and Hilgers (2014),Lévesque et al. (2010), Sylvestre et al. (2006), Jones and Fowler (2016)).

Immortal time refers to the precise span of time in which follow-up observation has already started, but treatment has not yet been assigned. Note that by design, a subject that would be exposed to the therapy in the future must remain event-free for an immortal time to be classified as treated.

This leads to the fact that for treated individuals, the event cannot occur within the specified period, resulting in a biased estimate of treatment efficacy. Immortal time bias cannot occur in RCTs, as follow-up begins at the same time as eligibility criteria are met and treatment is assigned.

The scheme in Figure 1.2 illustrates the various situations in which an incorrect adjustment for immortal time can generate bias.

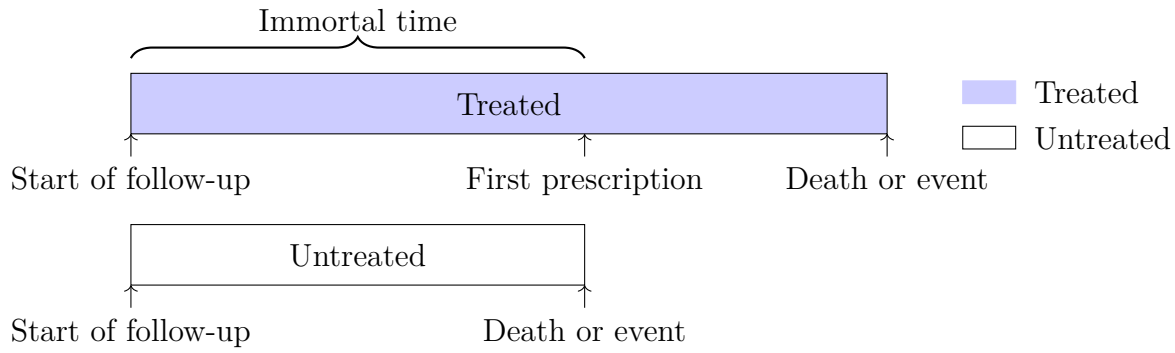


Figure 1.2: Visual representation of immortal time bias

Note that immortal time bias typically leads to an overestimation of therapy efficacy, posing a real threat of drawing incorrect conclusions from the data.

The presence of these biases, as well as all the other biases related to the three categories outlined above, makes the analysis of observational data particularly challenging,

necessitating the use of proper methodologies to address them.

1.2.1.3 Target trial emulation

Target trial emulation (TTE) (Hernán and Robins (2016), Gomes et al. (2022)), is a technique designed to estimate the effect of a particular treatment using observational data. It works by specifying and trying to reproduce (emulate) a specific RCT, that is, the one we would use to estimate the treatment effect of interest. Starting from observational data, TTE explicitly characterizes and emulates the target trial of interest following eight principal components (steps): *Eligibility criteria, Treatment strategies, Outcomes, Assignment procedures, Time zero, Estimands, Analysis plan.*

In the first three components (*Eligibility criteria, Treatment strategies*, and *Outcomes*), exclusion and inclusion criteria, treatment strategies, and primary and secondary outcomes are defined, respectively. They must reproduce the ones of the hypothetical trial we want to mimic. These steps are concretely realized by considering only people consistent with the targeted eligibility criteria, treatment strategies, and outcomes, respectively.

Assignment procedures is a crucial component, and it consists of retrieving the conditional exchangeability ensured in RCTs, as this property avoids different forms of selection bias and confounding. Ensuring conditional exchangeability also translates into addressing the bias related to the population imbalance described above. Different methods can be employed in this step (Rubin and Imbens (2015), Hernán and Robins (2016), Kurz (2022)):

- Matching methods: these methods work by matching each treated individual with one or more control individuals similar to them to construct balanced populations. The concept of similarity is usually associated with notions of distance between individuals' covariates (e.g. Mahalanobis matching) or the probability of individuals being treated given their characteristics (e.g. propensity score matching) (Cochran and Rubin (1973), Rubin (1973), Stuart (2010))

- Inverse probability weighting (IPW): this method assigns a weight to each individual in the data to obtain an unbiased estimate of the treatment effect. The weights assigned to each individual are $w_i = 1/e(\mathbf{X})$ if the individual i belongs to the treatment group, and $w_i = 1/(1 - e(\mathbf{X}))$ if the individual i belongs to the control group. Here, $e(\mathbf{X})$ is the propensity score, defined as $e(\mathbf{X}) = pr(Z = 1|\mathbf{X})$, i.e. the probability of being treated given the individual's covariates' profile $\boldsymbol{X}$. (Hernán and Robins (2020), Austin and Stuart (2015));

- Regression methods: these methods work by first fitting a regression on the conditional distribution of the outcome given the exposure variable and the confounders $Y|Z, \boldsymbol{X}$, and then averaging the predicted outcomes over the distribution of the confounders to estimate the relevant treatment effect (Hernán and Robins (2020), Kurz (2022));

- Augmented inverse probability weighting: this method works in two steps. First, it computes propensity scores. Second, it fits two outcome regression models (one for each group). Finally, each outcome is weighted with the propensity score of the first step to produce a weighted average of the two models (IPW and regression). This leads to a doubly robust estimator where just one of the propensity or outcome models needs to be correctly specified to obtain an unbiased estimate (Kurz (2022), Glynn and Quinn (2010)).

Note that there are many other interesting methods, such as g-estimation, machine learning methods and AI related methods (Hernán and Robins (2020), Wager and Athey (2018), Pearl (2009), Yao et al. (2021)), which are typically used to account for time-varying confounding and exposure, and on which we do not focus in this work, but that also play a very important role in HTA.

Another relevant component of the TTE process is the *Time zero* definition. In RCTs, time zero is defined as the time when eligibility criteria are met, treatment is assigned,

and follow-up begins. In observational data, these elements can be non-simultaneous. As we already mentioned, the misalignment of these three elements can cause different forms of bias, among which, one of the most important is *immortal time bias* (Tamm and Hilgers (2014),Lévesque et al. (2010), Sylvestre et al. (2006), Jones and Fowler (2016)). In this context, different approaches may be considered to properly define time zero (Karim et al. (2016), Wang et al. (2022)). For example, in studies where no therapy is administered to the control arm, we can either match pretreatment person-time between the two populations to make the start of follow-up coincide or rely on a particular form of nested trials.

The last two components of the TTE process are related to the analysis phase. In the *Estimands* stage, the choice between intention-to-treat or per-protocol analysis must be made. Once this choice is made in the *Analysis Plan* step, the methods to be applied for estimating the quantities of interest are defined. Standard methods applied at this stage are matching, inverse probability weighting, regression methods, and g-methods.

To summarize, the goal of TTE is to estimate the effect of a particular treatment using observational data, by emulating a targeted RCT starting from observational data, defining all relevant elements (eligibility criteria, treatment strategies, etc.), while adjusting for bias using various methodologies.

1.2.1.4 Chapter 2: Inverse target trial emulation and Observational data simulation

There are various challenges associated with applying TTE to inform decisions and estimate causal effects. In particular, there is substantial methodological interest in evaluating the different methods applied in the relevant components of TTE to analyse observational data. One key aspect of evaluating these methods is performing a *simulation study* that benchmarks the performance of the specific method against relevant alternative methods Morris et al. (2019). Simulation studies are computer experiments that create data

through pseudo-random sampling from known probability distributions before analyzing these data to determine key properties of the analysis methods Morris et al. (2019). Selecting a method to generate realistic data is, therefore, a key element of a simulation study. If the data-generating mechanism is inaccurate or unrealistic, then the study results may lack of validity, particularly when researchers have to analyze complex observational data. Therefore, in Chapter 2, we develop a general purpose method for generating realistic observational data.

Our methodology is based on having access to individual patient data from an initial Randomized Clinical Trial (RCT). From this, we define different ways of generating observational data using the information in the RCT dataset and different assumptions about the biases observed in the observational data. Once we generate observational data across a range of mechanisms, we can test a variety of causal inference methods to infer the relevant estimands. Finally, we can select the optimal causal inference method by determining which method is best able to recreate the estimated effect of interest from the original RCT. Note that the initial RCT is important not only to provide the true value of the estimand but also to simulate realistic relationships in the observational data. This is because it allows us to understand how the covariates are related to each other and the outcome of interest, which is the preliminary step of our data-generating process.

As we have mentioned in the previous section, in the causal inference literature, the target trial emulation (TTE) methodology (Hernán et al. (2016), Fu et al. (2021)) has been defined to link observational data to a hypothetical RCT. In this framework, TTE aims to reduce or eliminate bias by creating a link between a hypothetical RCT that could be used to answer the research question and the available observational data (García-Albéniz et al. (2017)). Drawing on this idea, our methodology to simulate observational data uses data from an RCT to clarify the population and relationship between the covariates before generating the observational data. In this sense, we were motivated by the definition and implementation of TTE and we have named our methodology *inverse target trial*

emulation (ITTE). Thus, in Chapter 2 we introduce inverse target trial emulation, a methodology to generate realistic observational data using available individual patient data from an RCT. We will discuss the steps of our proposed methodology before applying it to a range of realistic and relevant scenarios.

Moreover, the novel methodology can also play a crucial role in conducting various forms of probabilistic analysis in HTA. Suppose, for example, a scenario in which we plan to collect additional observational data to inform our decision model. In this context, we need to be able to simulate realistic observational data to determine in advance the types of data we can collect, thereby evaluating the amount of information they can provide to better estimate the relevant parameters in the economic model. We will deepen this application in the following chapters.

1.2.2 Randomized Controlled Trials

As we already mentioned, RCTs are typically considered the gold standard for informing decisions in HTA (Hariton and Locascio (2018)) because, if properly designed, they can avoid various forms of confounding, selection, and information bias §1.2.1. Thanks to the random allocation of patients to treatment or control groups, they can guarantee a balance between populations, while an accurate design can guarantee coherent and proper measurements and definitions. For these reasons, RCTs are particularly relevant when researchers have to evaluate the cost-effectiveness and safety of a novel treatment.

In many experiments related to RCTs, researchers consider the entire cohort of individuals to calculate the efficacy of a given treatment. However, in certain situations, it can happen that the treatment effect is heterogeneous among individuals with different characteristics. In these cases, the ‘one-size-fits-all‘ treatment recommendation should be discouraged as important information necessary to ensure the cost-effectiveness of the treatment of interest would be lost, although equity or other considerations may still justify such an approach in some contexts. Nevertheless, governments and other health-

related payers have recognized that collecting additional evidence and properly assessing heterogeneity can be highly beneficial in restricting public subsidies to individuals for whom it is cost-effective in terms of net benefit (Coyle et al. (2003)).

1.2.2.1 Chapter 5: Bayesian Subpopulation treatment effect pattern plot (B-STEPP)

There are various methods to calculate heterogeneity of treatment effect (HTE) between patients; these approaches measure the variation in effect across different levels of patients' characteristics (covariates) (Rothwell (2005), Kent et al. (2020), Künzel et al. (2019)).

Subgroup analysis is probably the most common approach (Rothwell (2005)). Patients are divided into distinct subsets based on their covariates or risk factors, and then an analysis is performed within each subgroup to detect heterogeneity in treatment efficacy.

However, this method presents different drawbacks. First, unless the sample size of each subgroup is sufficiently large, randomization cannot be assumed to hold within each subset a priori, as it is only ensured for the entire cohort or strata used in the randomization algorithm (Wang et al. (2010)). Second, if patients are categorized into disjoint subgroups, meaning that each distinct subset is characterized by a single summary index, the risk of losing important information on the effect of the baseline characteristic as a predictor of efficacy increases (Royston et al. (2006)). Third, analyzing disjoint subsets decreases the precision of the different relevant estimates as the sample size decreases (Lazar et al. (2016)). Finally, testing differences in efficacy across different subgroups, one variable at a time (e.g., old vs. young, male vs. female, smokers vs. non-smokers), increases the probability of chance findings; this problem is known as *multiplicity* (Wang et al. (2007)).

The problem of multiplicity is closely connected to another important aspect of subgroup analysis: it is crucial that this methodology is used to test for heterogeneity in the population related to a prespecified characteristic, as post hoc subgroup findings have

the potential to lead to incorrect conclusions. As stated in (Rothwell (2005)): *Selective reporting of post hoc subgroup observations, which are generated by the data rather than tested by them, is analogous to placing a bet on a horse after watching the race.*

To address some of the limitations above, another category of methods to investigate HTE in RCTs has been defined, that is: predictive HTE analysis (Kent et al. (2020)). Predictive HTE analysis objective is to determine individualized effectiveness differences across groups, by accounting for multiple relative characteristics simultaneously. It works in 2 different steps. First, one must set the variables and the model to define the different subgroups. Second, the effect estimation must be estimated across the stratified model described in the first step. This is concretely performed by specifying a regression equation on RCT data with as predictors both the covariates that characterize the subgroups (risk factors), the exposure variable, and covariates-exposure interaction terms.

Despite the precise and promising specification of this methodology, predictive HTE analysis still presents some potential risks of multiplicity. Moreover, in this scenario, we must specify the outcome model to define the regression equation, leading to a risk of model misspecification. In this regard, recent machine-learning approaches provide a promising solution, as they can flexibly learn complex functional forms and interaction structures directly from the data, substantially reducing dependence on model specification (Hahn et al. (2020), Henderson et al. (2020),Glynn et al. (2024),Hu et al. (2021)). Nevertheless, these methods typically require large sample sizes to achieve stable and reliable estimates of individualized treatment effects.

In (Bonetti and Gelber (2000)), the authors proposed the *Subpopulation treatment effect pattern plot* (STEPP), a nonparametric alternative to standard regression models, which typically require the specification of the relationship between the outcome and the covariates of interest. In this method, the authors divide the population with respect to a continuous covariate Z^*, defining subgroups that are no longer disjoint but rather overlap (Yip et al. (2016), Bonetti et al. (2009)).

In this context, cutpoints that generate subgroups are chosen based on two parameters, r_1 and r_2, with $r_1 < r_2 < n$. The former number, r_1, provides the (approximate) extent of overlap between the sub-populations in terms of number of units in common by each pair of consecutive sub-populations. The latter parameter, r_2, indicates the (approximate) sample size of each subpopulation. In their work, the authors proposed the an algorithm to divide the populations into subgroups with the prespecification of r_1 and r_2, we will describe the relevant algorithm in §5.1.

To give an intuitive representation, Figure 1.3 reports the generated structure of overlapping subgroups with the algorithm above.

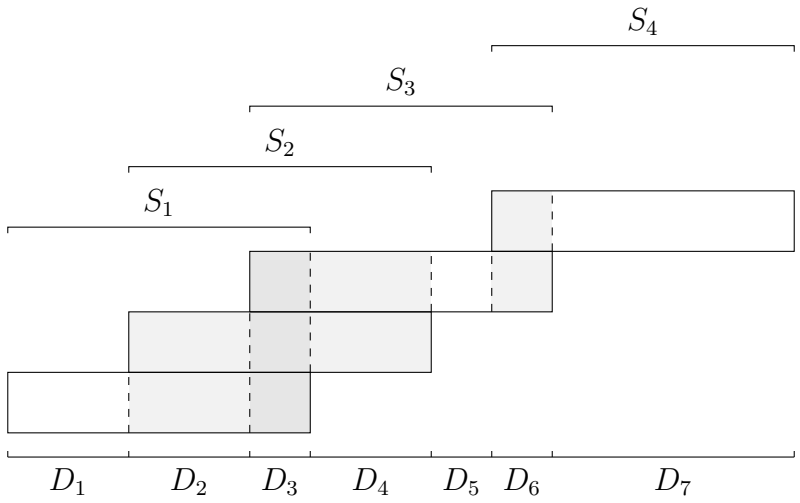


Figure 1.3: Illustration of four overlapping subpopulations, labeled $S_1, \ldots, S_4$ and the corresponding distinct subpopulations, labeled $D_1, \ldots, D_7$. The shaded regions indicate overlap

After the population is divided into subgroups, a chosen estimate of treatment effect is computed within each subset. Since subgroups are not disjoint, it is even more necessary to consider the dependence between the relevant estimates. To this aim, the authors compute the asymptotic joint distribution of the subgroup efficacy and use it for inference and hypothesis testing.

Note that considering overlapping subsets helps in reducing the problem related to loss of precision in standard subgroup analysis, as overlapping subgroups have a higher sample

size compared to disjoint subsets. In addition, STEPP does not assume an arbitrary specification of the subgroups; instead, the cutpoints are determined to ensure a specific sample size (or number of events) for each subgroup. Finally, STEPP assumes that it can divide the population based on a continuous covariate, thereby not losing relevant information on the effect of the baseline characteristic as a predictor of efficacy. STEPP has been applied in various realistic scenarios, yielding promising results in both inference and interpretability.

In Chapter 5, we propose a Bayesian version of STEPP, demonstrating that Bayesian statistics offer a highly flexible and natural approach to modeling dependence between efficacy estimates of different subpopulations. In addition, by relying on MCMC methods (Robert et al. (1999), Gilks et al. (1995)), we will demonstrate that a Bayesian approach enables us to make inferences assuming different flexible parametric models without relying on asymptotic distributions. Our first findings suggest that the Bayesian approach shows concrete computational time advantages. We will apply the novel method to two different real studies, the first with a binary outcome and the second in the survival setting. In both scenarios, we will sample from the joint posterior distribution of the subpopulations' outcome parameters for both treatment and control groups and construct the related credible intervals. In addition, we will sample from the posterior distributions of different relevant statistics and compute the credible intervals to verify the hypothesis of no difference in treatment effect across the subpopulations.

1.2.3 Missing data in evidence for health economic modeling

A particularly challenging phenomenon in health evidence analysis is the one related to missing data, which is common to all the different types of evidence collected to inform health economic models (Rubin (1976), Little and Rubin (2019),Gabrio et al. (2017)). Moreover, it is present in both RCTs and observational studies, particularly in health-related scenarios, and it can be due to various factors (e.g., patients lost to follow-up,

partially completed surveys, or incomplete medical records) (Marino et al. (2021)). The first step when we encounter missing data is to consider the possible reasons behind the missingness. In other words, we must make some assumptions about the underlying missing data mechanism.

We assume to have individual-level data coming from a health-related study (experimental or observational). In particular, we consider a sample of $i = 1, \ldots, n$ individuals, each one with K covariates $\boldsymbol{X} = (X_1, \ldots, X_K)$, outcome Y and exposure variable Z. In this context, we assume that while covariates and exposure variables are fully observed, the outcome can be missing; that is, outcomes are partially observed. Finally, we denote as π_i the missingness indicator for individual i; in particular, $\pi_i = 1$ if Y_i is missing and $\pi_i = 0$ if Y_i is observed.

In this context, there are different types of missing mechanisms a researcher can face while analyzing data (Schafer and Graham (2002),Faria et al. (2014)):

- **Missing completely at random (MCAR)**: when the probability of missingness does not depend on either the covariates $\boldsymbol{X}$ or the outcome Y, or both. In other words, π_i is independent of both Y_i, $\boldsymbol{X}_i$ and Z_i. This assumption is often unrealistic in real-world scenarios, where dropout from a study, loss to follow-up, partially completed surveys, incomplete medical records, and other reasons for missingness are often associated with certain patient characteristics.

- **Missing at random (MAR)**: when the missing generating mechanism depends on the covariates $\boldsymbol{X}$ but not on the outcome Y, or more in general it depends only on fully observed characteristics. In this context, it translates to π_i being independent of Y_i, but dependent on $\boldsymbol{X}_i$ and Z_i. For example, suppose that patients are more likely to drop out of a study if they are older because of difficulties in reaching the specific health center. In this case, π_i has a positive correlation with age.

- **Missing not at random (MNAR)** (Nonignorable missingness): when the missing

generating mechanism depends on both the covariates $\boldsymbol{X}$ and the outcome Y, in general it depends on both the fully and partially observed characteristics. In this context, it translates to π_i being dependent of both Y_i, $\boldsymbol{X}_i$ and Z_i. Consider the previous example again. Suppose that patients are more likely to drop out of a study if they are older due to difficulties in reaching the specific health center, but also if the therapy does not have the desired effect after the first session. In this case, π_i has a positive correlation with both age, and Y_i, the (if missing) unobserved outcome after the first session of therapy.

Note that the actual underlying missing mechanism cannot be identified from the observed data, which requires that an assumption must be made based on previous knowledge.

There are different methods that can be applied to analyze data affected by missingness:

- Complete case analysis (CCA): this method only considers fully observed data to make inferences and obtain relevant estimates. This potentially leads to incorrect estimates due to the disregard of relevant dependencies between π_i and individuals' characteristics, as well as the loss of power resulting from considering a subset of data with a smaller sample size.

- Single imputation: in this method, missing values are imputed with a single prediction. The latter can be obtained by taking the conditional mean or by considering other relevant values that can fit the missing observation (Buck (1960), Twisk and de Vente (2002), Shao and Zhong (2003)). The primary source of bias in these methods is that they do not account for the uncertainty in estimating missing data, failing to properly consider the variability of different imputation methods for the value of interest.

- Inverse probability weighting: this method starts with CCA and adjusts for bias estimates by assigning a weight to each observation, which is equal to the inverse of the probability of being observed, estimated assuming MAR using regression models or other relevant methods. Even if this method is preferable to standard CCA, it strongly relies on the correct specification of the weights' estimation model and moreover, it discards possible useful information for estimation in the partially observed individuals (Robins et al. (1994), Seaman and White (2013));

- Multiple imputation: this method involves imputing missing observations multiple times using different techniques to estimate the conditional predictive distribution of the missing values given the observed characteristics. After generating different complete imputed datasets, they are combined, using the Rubin rule or other techniques, to obtain relevant estimates of the missing values (Rubin (1987)).

Many other methods exist to account for missing data, each one with its own advantages and disadvantages. For a more complete list, see (Gabrio et al. (2017)). Note that the majority of the methods above work by assuming a MAR missing mechanism. Therefore, in realistic studies when we are not sure about the MAR assumption, we might perform different forms of sensitivity analysis to understand how much the relevant estimates change for different MNAR models, or, we might directly try to model the MNAR missing mechanism explicitly (Molenberghs et al. (2014), Gabrio et al. (2019)).

1.3 Value of Information with generalized data collection schemes

In the last part of §1.1.1, we mentioned the possibility, within a Bayesian decision framework, of collecting further evidence to inform our economic model, thereby reducing the uncertainty associated with related parameters. This possibility is realized by Value of

Information (VoI), a methodology that encompasses both sensitivity analysis, research prioritization, and study design.

VoI basically answers these fundamental questions: *"On which part of the model would further information be most valuable?"*, and *"What is the value of collecting a specific number of observations in a study with a particular design?"* (Jackson et al. (2022)).

1.3.1 Value of Information

Suppose we are in a Bayesian decision framework and we have a model that incorporates different sources of information. One can be interested in determining which parameters of the model drive the most uncertainty in both the estimates and the decision to be made. Moreover, one may want to know what the expected reduction in uncertainty is if new data are collected and to understand the economic value of reducing uncertainty to make better decisions.

To properly define VoI measures we assume to be in a Bayesian decision theoretic framework and consider a decision model as the one introduced in §1.1.1, with a set of parameters $\boldsymbol{\theta}$ and expected net benefit $\mathbb{E}(\text{NB}_d(\boldsymbol{\theta}))$ with $d = 1, \ldots, D$ possible decisions. The uncertainty contained in the model is represented by a probability distribution $p(\boldsymbol{\theta})$, which can be viewed as either a prior distribution or a posterior distribution after incorporating all available knowledge. In this framework, the objective of the decision model is to choose the action d that maximizes the expected utility (net benefit) $\mathbb{E}(\text{NB}_d(\boldsymbol{\theta}))$.

In the context above, a VoI analysis quantifies the expected benefit, i.e. the reduction in expected loss that we have by collecting new information to reduce uncertainty in the model. The following are the main VoI measures:

- The **Expected Value of Perfect Information** (EVPI) is the expected net benefit gain by having perfect information on $\boldsymbol{\theta}$, i.e., knowing precisely the value of $\boldsymbol{\theta}$:

$$\text{EVPI} = \mathbb{E}_{\boldsymbol{\theta}} \max_d \{\text{NB}_d(\boldsymbol{\theta})\} - \max_d \{\mathbb{E}_{\boldsymbol{\theta}}(\text{NB}_d(\boldsymbol{\theta}))\} \tag{1.3}$$

- The **Expected Value of Partial Perfect Information** (EVPPI) of a subset of parameters $\boldsymbol{\phi} \subset \boldsymbol{\theta}$ is the expected net benefit gain in learning $\boldsymbol{\phi}$ precisely (Heath et al. (2017))

$$\text{EVPPI}_{\phi} = \mathbb{E}_{\boldsymbol{\phi}} \max_d \{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\phi}} \text{NB}_d(\boldsymbol{\theta})\} - \max_d \{\mathbb{E}_{\boldsymbol{\theta}}(\text{NB}_d(\boldsymbol{\theta}))\} \tag{1.4}$$

- The **Expected Value of Sample Information** (EVSI) (Raiffa and Schlaifer (2000), Ades et al. (2004)) is the expected net benefit gain in collecting an additional sample $\mathbf{y}$

$$\text{EVSI}_{\boldsymbol{y}} = \mathbb{E}_{\boldsymbol{Y}} \max_d \{\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}} \text{NB}_d(\boldsymbol{\theta})\} - \max_d \{\mathbb{E}_{\boldsymbol{\theta}}(\text{NB}_d(\boldsymbol{\theta}))\} \tag{1.5}$$

The rationale behind the definition of the EVSI can be illustrated by composing the expression (1.5) in different steps.

At the beginning, before collecting new data, the best treatment is the one that maximises the expected net benefit,

$$\max_d \mathbb{E}_{\boldsymbol{\theta}} \left(\text{NB}_d(\boldsymbol{\theta})\right) \quad d = 1, \ldots, D$$

The next step is to collect new data $\boldsymbol{y}$ from the Likelihood

$$\boldsymbol{Y}|\boldsymbol{\theta} \sim F(\boldsymbol{\theta})$$

Once $\boldsymbol{y}$ are collected, the best decision would be the one that maximizes the conditional expectation of $\boldsymbol{\theta}$ given the data we collect

$$\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y}} \left(\mathrm{NB}_d(\boldsymbol{\theta})\right)$$

Since the data are not collected at the time we plan the study, we must average over all the possible data we may collect in the future and subtract the expected value of the initial optimal decision to obtain the definition of the EVSI:

$$\mathrm{EVSI} = \mathbb{E}_{\boldsymbol{Y}} \left[\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y}} \left(\mathrm{NB}_d(\boldsymbol{\theta})\right)\right] - \max_d \mathbb{E}_{\boldsymbol{\theta}} \left(\mathrm{NB}_d(\boldsymbol{\theta})\right) \tag{1.6}$$

1.3.2 EVSI computation

An interesting challenge involves finding efficient computational methods to calculate the key VoI measures just presented. Since our work focuses on EVSI, we have to describe the principal methods for computing it.

The calculation of EVSI is the most challenging within VoI measures due to both the task of generating realistic artificial data and the estimation of the posterior distribution $\boldsymbol{\theta}|\boldsymbol{Y}$ or $\boldsymbol{\phi}|\boldsymbol{Y}$, depending on the parameters the data $\boldsymbol{Y}$ inform. For this reason, EVSI has been applied only to fairly simple data collection exercises to date (Heath et al. (2022)).

The most commonly used methods to compute EVSI are the following.

- Straightforward **brute-force nested Monte Carlo** method (Ades et al. (2004)):

$$\hat{\mathrm{EVSI}} = \frac{1}{K}\sum_{k=1}^{K} \max_d \frac{1}{J}\sum_{j=1}^{J} \mathrm{NB}_d(\theta^{(j,k)}) - \max_d \frac{1}{N}\sum_{n=1}^{N} \mathrm{NB}_d(\theta^{(n)})$$

  where $\boldsymbol{\theta}^{(j,k)}$ are samples drawn from the posterior distribution of $\boldsymbol{\theta} \mid \mathbf{Y}^{(k)}$, and $\mathbf{Y}^{(k)}$ are generated by first sampling $\boldsymbol{\theta}^{(k)}$ from $p(\boldsymbol{\theta})$ and then $\mathbf{Y}^{(k)}$ from $p(\mathbf{Y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}^{(k)})$. As one can immediately notice, the brute force approach has a high computational burden;

- **Moment matching method**: this method is devoted to estimating the distribu-

tion of the preposterior mean

$$\mu_t^{(\boldsymbol{Y})} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\text{NB}_t(\boldsymbol{\theta}))$$

To do this, the prior distribution of the net benefit $\text{NB}_t(\boldsymbol{\theta})$is rescaled to match the mean and variance of the preposterior mean computed separately using MC approximation methods. Once we have estimated the distribution of $\mu_t^{(\boldsymbol{y})}$, EVSI is estimated simply by using MC integration. Note that the key assumption here is that the distribution of the preposterior mean has a similar form to that of the net benefit $\text{NB}_t(\boldsymbol{\theta})$. This assumption holds if the data $\boldsymbol{Y}$ informs the entire set of parameters $\boldsymbol{\theta}$, which is not very realistic in many settings. When data inform only a subset $\boldsymbol{\phi}$ of parameters, then the distribution we take as the baseline to approximate the preposterior mean is that of $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\phi}}(\text{NB}_t(\boldsymbol{\theta}))$ (Heath et al. (2018), Heath et al. (2019)).

- **Nonparametric regression**: In this approach, multiple nonparametric regression equations (one for each possible treatment) are fitted using nonparametric regression models to compute the preposterior mean. The two most common classes of regression models are *generalized additive models* (GAMs) (Wood (2017)), and *gaussian processes* (Seeger (2004)). Usually, GAMs are used when the number of predictors is low (usually less than 5 (Heath et al. (2017))), as computing the interaction between multiple predictors in their relation with the outcome can be infeasible as the number of predictors increases. Conversely, GP is less problematic in accounting for the complexity introduced by a higher number of predictors, even if for a low number of the latter, GAMs remain the most computationally efficient choice; for a deeper comparison of these and other possible models, see (Heath et al. (2024)). Note that here we refer to predictors as the parameters for which we want to compute the joint EVSI, that is, the parameters we expect to be informed by the data

we collect. In this work, we will focus mainly on GAM to estimate the preposterior mean in nonparametric regression equations (Strong et al. (2015)):

$$\mathrm{NB}_d(\boldsymbol{\theta}) = g_d(T(\boldsymbol{Y})) + \epsilon$$

where $\boldsymbol{\theta}$ is a sample of the joint distribution of the input parameters obtained by performing a probabilistic analysis (PA), $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K)$ are the sampled data, with $\boldsymbol{Y}_i$ simulated from the predictive distribution $\boldsymbol{Y}|\theta^{(i)}$, with $\theta^{(i)}$ i-th sample of PA simulation (see Heath et al. (2022) for details). Finally, T is some low-dimensional summary statistic of the data. Ideally, T should incorporate all the information in the original data; for this reason, it is usually suggested to choose sufficient statistics of the parameters we assume are informed by the data. Other reliable options, when sufficient statistics are not obtainable, are typically sample estimators or maximum Likelihood estimators (MLE) of the parameter(s) we expect to inform collecting new data (Li et al. (2024)). Suppose, for instance, $(\phi^{(1)}, \phi^{(2)}) \subseteq \boldsymbol{\theta}$ are the parameters we expect to inform by collecting new data $\boldsymbol{Y}$. In this case, we could have $\boldsymbol{T}(\boldsymbol{Y}) = (T^{(1)}(\boldsymbol{Y}), T^{(2)}(\boldsymbol{Y}))$ where $T^{(i)}(\boldsymbol{Y})$ is a sample or maximum Likelihood estimator of $(\phi^{(1)}, \phi^{(2)})$. Once an estimate of the preposterior mean is obtained, EVSI is computed through classical Monte Carlo integration;

There are several other methods to compute EVSI, such as Gaussian approximation (Jalal and Alarid-Escudero (2018), Li et al. (2024)), Importance sampling (Menzies (2016), Heath et al. (2020)), and others. For a review and a comparison of the different methods, see (Kunst et al. (2020)).

1.3.3 EVSI and complex data collection mechanism

EVSI has been applied only to idealized and simple data collection exercises to date (Heath et al. (2022)). In most cases, it has been used to understand the value of idealized RCT

data, meaning that it measures the value of collecting additional idealized RCT data to reduce uncertainty in the underlying health economic model. However, we already highlighted that RCTs can suffer from high costs and difficulties in feasibility in different contexts. Moreover, limiting the usage of VoI techniques to idealized data collection schemes does not allow researchers to compute the value of additional data in more realistic contexts. For these reasons, we would like to explore a range of alternative study types for use in our EVSI calculations.

1.3.3.1 Chapter 3: EVSI and missing data

As we already mentioned in §1.2.3, missing data are quite common in health economics studies. To our knowledge, no methodology has yet been proposed to compute EVSI with a realistic missing mechanism that affects the data one plans to collect.

In Chapter 3, we define a methodology for computing EVSI with a more complex data-generating mechanism. In particular, we compute EVSI in the scenario in which the data we plan to collect in the future have different missing observations. This is particularly relevant since, in realistic applications, this is usually the case. Understanding the value of data with missingness can also be helpful in determining the optimal sample size to collect, thereby maximizing the value of future data in reducing uncertainty around the relevant economic model parameters. Moreover, as in the standard procedure, understanding how to apply EVSI in this context would be helpful in designing experiments when planning to collect data that suffers from missingness (research prioritization).

As we defined in §1.2.3, there are different types of missing mechanisms a researcher can face while analyzing data, namely *missing completely at random, missing at random, missing not at random*. Until now, EVSI on missing data has been computed only with data affected by an MCAR missing mechanism (Heath et al. (2022)), which is unrealistic in real practice. For this reason, we aim to develop a more comprehensive methodology that can manage more complex and realistic missing data mechanisms. In Chapter 3,

we will define various methods for generating data with the above types of missingness and then focus on the last two, which are the most realistic in practical applications to develop a methodology for computing EVSI in this scenario. Finally, we apply this novel methodology first to synthetic data, generated assuming a Normal conjugate model, and then to the more realistic example introduced in Section §1.1.2.

1.3.3.2 Chapter 4: EVSI and observational data

Introducing a methodology to compute EVSI when planning to collect data with missing values is crucial, as it enables us to model more realistic contexts, rather than relying on idealized data collection exercises. Despite this, we would like to explore not only more realistic scenarios but also a wider range of alternative study types for use in our EVSI calculations. Therefore, in Chapter 4, we define a novel methodology for applying EVSI when planning to collect observational data. To achieve this, we first generate observational data, as described in Chapter 2, with a focus on the bias related to population imbalance. We then apply an analogous version of the methodology introduced in Chapter 3 to compute EVSI with observational data.

Understanding how to perform EVSI in face of different types of observational data, with varying associated biases, can be particularly useful, as it would allow us to utilize such a powerful instrument in contexts where we are not as concerned about the accessibility and costs of data. Moreover, as with the application of RCTs, understanding how to apply EVSI in this novel context would be highly beneficial in designing experiments when planning to collect observational data (research prioritization).

Again, we apply this novel methodology first to synthetic data, generated assuming a Normal conjugate model, and then to the more realistic example introduced in Section §1.1.2.

References

Ades, A., Lu, G., and Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling. *Medical decision making*, 24(2):207–227.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679.

Baracskay, D. (1998). Cost-benefit analysis: Concepts and practice.

Bonetti, M. and Gelber, R. D. (2000). A graphical method to assess treatment–covariate interactions using the cox model on subsets of the data. *Statistics in medicine*, 19(19):2595–2609.

Bonetti, M., Zahrieh, D., Cole, B. F., and Gelber, R. D. (2009). A small sample study of the stepp approach to assessing treatment–covariate interactions in survival data. *Statistics in medicine*, 28(8):1255–1268.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A., Karnon, J., Sculpher, M. J., Paltiel, A. D., Force, I.-S. M. G. R. P. T., et al. (2012). Model parameter estimation and uncertainty: a report of the ispor-smdm modeling good research practices task force-6. *Value in Health*, 15(6):835–842.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(2):302–306.

Carlson, M. D. and Morrison, R. S. (2009). Study design, precision, and validity in observational studies. *Journal of palliative medicine*, 12(1):77–82.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P. C., and Sculpher, M. (2015). Methods for the estimation of the national institute for health and care excellence cost-effectiveness threshold. *Health Technology Assessment (Winchester, England)*, 19(14):1.

Cochran, W. G. and Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*, volume 1195. Houghton Mifflin Boston, MA.

Coyle, D., Buxton, M. J., and O'Brien, B. J. (2003). Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. *Health economics*, 12(5):421–427.

Faria, R., Gomes, M., Epstein, D., and White, I. R. (2014). A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *Pharmacoeconomics*, 32(12):1157–1170.

Fenwick, E., Claxton, K., Sculpher, M., and Briggs, A. (2000). Improving the efficiency and relevance of health technology assessment: the role of iterative decision analytic modelling. *DISCUSSION PAPER-UNIVERSITY OF YORK CENTRE FOR HEALTH ECONOMICS*.

Fu, E. L., Evans, M., Clase, C. M., Tomlinson, L. A., van Diepen, M., Dekker, F. W., and Carrero, J. J. (2021). Stopping renin-angiotensin system inhibitors in patients

with advanced ckd and risk of adverse outcomes: a nationwide study. *Journal of the American Society of Nephrology*, 32(2):424–435.

Gabrio, A., Mason, A. J., and Baio, G. (2017). Handling missing data in within-trial cost-effectiveness analysis: a review with future recommendations. *PharmacoEconomics-open*, 1(2):79–97.

Gabrio, A., Mason, A. J., and Baio, G. (2019). A full bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. *Statistics in medicine*, 38(8):1399–1420.

García-Albéniz, X., Hsu, J., and Hernán, M. A. (2017). The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *European journal of epidemiology*, 32:495–500.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

Glynn, D., Giardina, J., Hatamyar, J., Pandya, A., Soares, M., and Kreif, N. (2024). Integrating decision modeling and machine learning to inform treatment stratification. *Health Economics*, 33(8):1772–1792.

Gomes, M., Latimer, N., Soares, M., Dias, S., Baio, G., Freemantle, N., Dawoud, D., Wailoo, A., and Grieve, R. (2022). Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *Pharmacoeconomics*, 40(6):577–586.

Grimes, D. A. and Schulz, K. F. (2002). Bias and causal associations in observational research. *The lancet*, 359(9302):248–252.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

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

Heath, A., Kunst, N., and Jackson, C. (2024). *Value of information for healthcare decision-making*. CRC Press.

Heath, A., Kunst, N., Jackson, C., Strong, M., Alarid-Escudero, F., Goldhaber-Fiebert, J. D., Baio, G., Menzies, N. A., and Jalal, H. (2020). Calculating the expected value of sample information in practice: considerations from 3 case studies. *Medical Decision Making*, 40(3):314–326.

Heath, A., Manolopoulou, I., and Baio, G. (2017). A review of methods for analysis of the expected value of information. *Medical decision making*, 37(7):747–758.

Heath, A., Manolopoulou, I., and Baio, G. (2018). Efficient monte carlo estimation of the expected value of sample information using moment matching. *Medical Decision Making*, 38(2):163–173.

Heath, A., Manolopoulou, I., and Baio, G. (2019). Estimating the expected value of sample information across different sample sizes using moment matching and nonlinear regression. *Medical Decision Making*, 39(4):347–359.

Heath, A., Strong, M., Glynn, D., Kunst, N., Welton, N. J., and Goldhaber-Fiebert, J. D. (2022). Simulating study data to support expected value of sample information calculations: a tutorial. *Medical Decision Making*, 42(2):143–155.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764.

Hernán, M. A. and Robins, J. M. (2020). Causal inference: What if.

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., and Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology*, 79:70–75.

Hu, L., Ji, J., and Li, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in medicine*, 40(21):4691–4713.

Hunink, M. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., and Glasziou, P. P. (2014). *Decision making in health and medicine: integrating evidence and values*. Cambridge university press.

Imbens, G. W. and Rubin, D. B. (2015). Assessing overlap in covariate distributions. *Causal inference in statistics, social, and biomedical sciences*.

Jackson, C. H., Baio, G., Heath, A., Strong, M., Welton, N. J., and Wilson, E. C. (2022). Value of information analysis in models to inform health policy. *Annual Review of Statistics and Its Application*, 9:95–118.

Jalal, H. and Alarid-Escudero, F. (2018). A gaussian approximation approach for value of information analysis. *Medical Decision Making*, 38(2):174–188.

Jones, M. and Fowler, R. (2016). Immortal time bias in observational studies of time-to-event outcomes. *Journal of critical care*, 36:195–199.

Karim, M. E., Gustafson, P., Petkau, J., Tremlett, H., Benefits, L.-T., of Beta-Interferon for Multiple Sclerosis (BeAMS) Study Group, A. E., Ehsanul Karim, M., Gustafson, P., Petkau, J., Tremlett, H., Shirani, A., et al. (2016). Comparison of statistical approaches for dealing with immortal time bias in drug effectiveness studies. *American journal of epidemiology*, 184(4):325–335.

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of internal medicine*, 172(1):35–45.

Kunst, N., Wilson, E. C., Glynn, D., Alarid-Escudero, F., Baio, G., Brennan, A., Fairley, M., Goldhaber-Fiebert, J. D., Jackson, C., Jalal, H., et al. (2020). Computing the expected value of sample information efficiently: practical guidance and recommendations for four model-based methods. *Value in Health*, 23(6):734–742.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167.

Lazar, A. A., Bonetti, M., Cole, B. F., Yip, W.-k., and Gelber, R. D. (2016). Identifying treatment effect heterogeneity in clinical trials using subpopulations of events: Stepp. *Clinical Trials*, 13(2):169–179.

Lévesque, L. E., Hanley, J. A., Kezouh, A., and Suissa, S. (2010). Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *Bmj*, 340.

Li, L., Jalal, H., and Heath, A. (2024). Estimating the evsi with gaussian approximations and spline-based series methods. *arXiv preprint arXiv:2401.17393*.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Marino, M., Lucas, J., Latour, E., and Heintzman, J. D. (2021). Missing data in primary care research: importance, implications and approaches. *Family practice*, 38(2):199–202.

McLachlan, G. J. (1999). Mahalanobis distance. *Resonance*, 4(6):20–26.

Menzies, N. A. (2016). An efficient estimator for the expected value of sample information. *Medical Decision Making*, 36(3):308–320.

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., and Ganiats, T. G. (2016). *Cost-effectiveness in health and medicine*. Oxford University Press.

O'Brien, B. J., Drummond, M. F., Labelle, R. J., and Willan, A. (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical care*, pages 150–163.

Parmigiani, G. and Inoue, L. (2009). *Decision theory: Principles and approaches*. John Wiley & Sons.

Pearl, J. (2009). Causal inference in statistics: An overview.

Raiffa, H. (1968). Decision analysis: Introductory lectures on choices under uncertainty.

Raiffa, H. and Schlaifer, R. (2000). *Applied statistical decision theory*. John Wiley & Sons.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141.

Rubin, D. and Imbens, G. (2015). Assessing overlap in covariate distributions. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, pages 309–336.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). Multiple imputation for survey nonresponse.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Seeger, M. (2004). Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106.

Shao, J. and Zhong, B. (2003). Last observation carry-forward and last observation analysis. *Statistics in medicine*, 22(15):2429–2441.

Stinnett, A. A. and Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical decision making*, 18(2_suppl):S68–S80.

Strong, M., Oakley, J. E., Brennan, A., and Breeze, P. (2015). Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Medical Decision Making*, 35(5):570–583.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Sylvestre, M.-P., Huszti, E., and Hanley, J. A. (2006). Do oscar winners live longer than less successful peers? a reanalysis of the evidence. *Annals of internal medicine*, 145(5):361–363.

Tamm, M. and Hilgers, R.-D. (2014). Chronological bias in randomized clinical trials arising from different types of unobserved time trends. *Methods of Information in Medicine*, 53(06):501–510.

Thokala, P., Ochalek, J., Leech, A. A., and Tong, T. (2018). Cost-effectiveness thresholds: the past, the present and the future. *Pharmacoeconomics*, 36(5):509–522.

Twisk, J. and de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of clinical epidemiology*, 55(4):329–337.

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., and Van Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child development*, 85(3):842–860.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Walker, S., Griffin, S., Asaria, M., Tsuchiya, A., and Sculpher, M. (2019). Striving for a societal perspective: a framework for economic evaluations when costs and effects fall on multiple sectors and decision makers. *Applied health economics and health policy*, 17(5):577–590.

Wang, J., Peduzzi, P., Wininger, M., and Ma, S. (2022). Statistical methods for accommodating immortal time: A selective review and comparison. *arXiv preprint arXiv:2202.02369*.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194.

Wang, S.-J., O'Neill, R. T., and Hung, H. J. (2010). Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. *Clinical Trials*, 7(5):525–536.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

Yip, W.-K., Bonetti, M., Cole, B. F., Barcella, W., Wang, X. V., Lazar, A., and Gelber, R. D. (2016). Subpopulation treatment effect pattern plot (stepp) analysis for continuous, binary, and count outcomes. *Clinical Trials*, 13(4):382–390.

Chapter 2

Inverse Target Trial Emulation

2.1 Introduction

Observational or real-world data, which are collected during routine practice, are increasingly used in health and medicine because they are widely available, relatively inexpensive to access and can be representative of the general population (Concato et al. (2017), Hilton Boon et al. (2022)). Despite this, there are challenges associated with using observational data to inform decisions and estimate causal effects, as they are often biased. Given these biases, it is critical to analyze observational data using appropriate methods to accurately draw conclusions (Rubin (1974)).

Thus, there is substantial methodological interest in developing novel methods to analyse observational data to adjust for these biases. One key aspect of evaluating these novel methods is performing a *simulation study* that benchmarks the performance of the novel method against relevant alternative methods Morris et al. (2019). Simulation studies are computer experiments that create data through pseudo-random sampling from known probability distributions before analyzing these data to determine key properties of the analysis methods Morris et al. (2019). Selecting a method to generate realistic data is, therefore, a key element of a simulation study. If the data-generating mechanism is inaccurate or unrealistic, then the study results may lack of validity, particularly when researchers have to analyze complex observational data. Therefore, in this chapter, we develop a general purpose method for generating realistic observational data.

It is crucial that this data-generating process covers different scenarios and assumptions on how the data arise. Typically, data can either be simulated from a known statistical model (Austin and Schuster (2016), Lunceford and Davidian (2004)) or researchers can use resampling with replacement from an available dataset (Bottigliengo et al. (2021), Desai et al. (2019)). Both these approaches have different limitations. For the first approach, it can be challenging to determine the true data-generating mechanism and estimand and the results can be highly model dependent. For the second approach,

the results may only be valid for that dataset and the number of questions that can be answered may be limited, for example, it may not be possible to assess bias due to population mis-specification. Thus, our approach combines elements from both these approaches.

Our methodology is based on having access to individual patient data from an initial Randomized Clinical Trial (RCT). From this, we define different ways of generating observational data using the information in the RCT dataset and different assumptions about the biases observed in the observational data. In this way, we avoid specifying the complete data-generating process and not do restrict our results to a specific bias-generation method. Once we generate observational data across a range of mechanisms, we can test a variety of causal inference methods to infer the relevant estimands. Finally, we can select the optimal causal inference method by determining which method is best able to recreate the estimated effect of interest from the original RCT. Note that the initial RCT is important not only to provide the true value of the estimand but also to simulate realistic relationships in the observational data. This is because it allows us to understand how the covariates are related to each other and the outcome of interest, which is the preliminary step of our data-generating process.

In the causal inference literature, the target trial emulation (TTE) methodology (Hernán et al. (2016), Fu et al. (2021)) has been defined to link observational data to a hypothetical RCT. In this framework, TTE aims to reduce or eliminate bias by creating a link between a hypothetical RCT that could be used to answer the research question and the available observational data (García-Albéniz et al. (2017)). Drawing on this idea, our methodology to simulate (not only emulate) observational data uses data from an RCT to clarify the population and relationship between the covariates before generating the observational data. In this sense, we were motivated by the definition and implementation of TTE and we have named our methodology *inverse target trial emulation* (ITTE). Note that, while "emulate" in TTE is referred to estimating causal effects using real-world data in a way that mimics a randomized trial without explicitly simulating RCT data;

in ITTE, we do not emulate or imitate data to draw conclusions but explicitly simulate them to generate individual level observational data in different ways and under different assumptions.

Thus, in this paper we introduce inverse target trial emulation, a methodology to generate realistic observational data using available individual patient data from an RCT. We will discuss the steps of our proposed methodology before applying it to a range of realistic and relevant scenarios. In section §2.2, we describe how to go from Target trial emulation to Inverse target trial emulation by inverting each step of TTE, adding bias inside the randomized data. In section §2.3, we define the different methods to perform ITTE. In section §2.4, we define the tests to evaluate the robustness and effectiveness of the different methods used to adjust for bias in observational data, enabling proper analysis. Note that the latter can be achieved by performing Inverse target trial emulation and Target trial emulation consecutively, and comparing the estimands of the starting RCT with those from the final target trial. In section §2.5, we present the results and discuss them.

2.2 Inverse Target Trial Emulation

To present our method for simulating realistic observational data from initial experimental data, we first introduce some additional notation.

2.2.1 Notation

We consider a dataset D_{rct} from a randomized controlled trial as individual-level data composed of a set of different covariates $\boldsymbol{X}$, the exposure variable Z, the outcome variable(s) $\boldsymbol{Y}$ and, if we are in a survival setting, the censoring variable C. Similarly, we introduce D_{obs} as individual-level data collected from an observational study. In this context, a general individual I_i represents the observation $(Y_i, \boldsymbol{X}_i, Z_i, C_i)$.

As we described in §1.2.1.3, the goal of Target trial emulation is, starting from D_{obs}, to emulate the related D_{rct}, defining all relevant elements (eligibility criteria, treatment strategies,...), while adjusting for bias using different methodologies.

Conversely, Inverse target trial emulation aims to exploit all relevant information contained in D_{rct} and to simulate (not only emulate) D_{obs}, adding different forms of bias and noise inside the data. It is essential to specify that while in Target trial emulation, we only **emulate** an idealized RCT (by adjusting the bias in the initial observational data) that we define and target at the beginning of the TTE process; in the ITTE methodology, we concretely **simulate** coherent non-experimental data.

2.2.2 From Target trial emulation to Inverse target trial emulation

To introduce ITTE, the idea is to exploit all the information in the initial RCT to simulate realistic observational data with different types of bias by revisiting and "inverting" all the components of TTE we introduced in §1.2.1.3. In particular, the aim is to reproduce the bias typically encountered in observational data by "inverting" the "assignment procedures" and "definition of time zero" components of a target trial emulation.

We now describe our novel framework, revisiting and inverting all the previous steps except the last two, which are mainly related to the analysis phase.

The three initial components of the TTE process that we review in this novel framework are: *Eligibility criteria, Treatment strategies*, and *Outcomes* (Hernán and Robins (2016)). To do this, we first need to collect all the relevant information from the RCT related to these components, namely: the characteristics of the patients (covariates) and their joint distribution, the treatment strategies related to the randomized study, and finally, the measured outcome and its (conditional) distribution. After we collect this information, we can generate a synthetic data set consistent with the initial RCT, as we

describe in §2.2.2.1.

Finally, we know that in realistic observational data, we usually discard a significant amount of information during a TTE process (e.g., patients not included in the eligibility criteria, irrelevant covariates, and measured outcomes). For this reason, we can include patients who do not meet the eligibility criteria or have treatment strategies and outcomes different from those of the randomized study. Another element of "noise" that we can add is missingness in the simulated dataset, as it is known that observational data tend to suffer from missingness more than experimental data.

The *Assignment procedures* component in this novel framework aims to replicate the various forms of bias observed in observational data in terms of population balance. As the confounders are asymptotically balanced in the RCT, our aim is to introduce bias to make the two arms unbalanced and reduce overlap. In this context, confounders may be either observed (measured covariates affecting both treatment and outcome) or unobserved (variables influencing both but not recorded in the dataset, or perhaps completely unknown). In the ITTE setting, we focus on inducing bias through the observed confounders only, as unobserved confounders cannot be explicitly modelled within the framework. We will introduce this in §2.2.2.2 while in the next section, we will define some methods that work under different assumptions to build an unbalanced observational dataset with varying degrees of overlap and varying degrees of knowledge about the treatment allocation mechanism.

The final step to invert in the novel framework is the *Time zero* component (Suissa (2008), Lévesque et al. (2010)). In this step to add time-related bias we first have to simulate the arrival of patients together with their grace period (immortal time), their survival outcome, and the censoring variable and then add bias by rescaling the survival outcome conditioning to the grace period or adding some dependency between the outcome variable and the arrival time. We detail this process in the next section.

We now deepen the general methodology to define ITTE, describing it in three blocks

above, presenting the main general steps, while in the next section we will provide a detailed description of the methods themselves. §2.2.2.1 is related to the *eligibility criteria, treatment strategies* and *results* steps and describes how we learn from the RCT. The second block in §2.2.2.2 is related to the *assignment procedure* component and explains how to unbalance observed confounders, and the last block in §2.2.2.3 is related to generating bias in the definition of *time zero*.

2.2.2.1 Extrapolation of information from RCT

What we are trying to achieve in this section is to simulate a realistic population consistent with the initial randomized trial D_{rct}. With "consistency", here, we mean that the simulated data must preserve the population structure of the RCT in terms of covariates distribution and correlation between covariates and the outcome distribution.

It is crucial to preserve the underlying structure of the RCT to both reproduce a realistic population that is coherent with the one that we might collect in future observational data and to allow for comparison between different TTE methods. Notice that, also in the TTE procedure, we maintain coherence between the starting D_{obs} and the target D_{rct}.

To do this, we have to extrapolate all relevant information from D_{rct}. The information we extrapolate from the RCT refers to 1) the distributions of covariates and their correlation, and 2) the conditional distribution $Y|\boldsymbol{X}, Z$ of the outcome given the exposure variable and the covariates. Note that we cannot learn anything about the conditional dependence of the exposure variable on the covariates, since patients are randomly assigned in an RCT.

Learning from data is essential in this context, as it enables us to reproduce all relevant dependencies and sample data with coherent and realistic behavior. Bayesian methods allow us to pursue this goal in a natural and flexible way; to this end, we introduce the Bayesian nested regression model below.

We start from a general RCT with a population of N individuals. For each individual,

the randomized data contain K covariates $\{X_1, X_2, .., X_K\}$, the exposure variable Z, the outcome variable Y and, in the case of a survival study, the censoring variable C. We factorize the joint distribution of the covariates and the outcome variable in a product of sequential conditional distributions. In particular, for X_1, we first define the marginal Likelihood distribution as:

$$X_1 \sim F_1(\boldsymbol{\Theta}_0)$$

and the functional form of the parameters $\boldsymbol{\Theta}_0$ itself as:

$$\boldsymbol{\Theta}_0 = g_0(\boldsymbol{\alpha}_{\mathbf{0}}^{(\mathbf{1})})$$

we then define a prior distribution for $\boldsymbol{\alpha}_{\mathbf{0}}^{(\mathbf{1})}$:

$$\boldsymbol{\alpha}_{\mathbf{0}}^{(\mathbf{1})} \sim G_1$$

Next, we define both the marginal Likelihood of the conditional distribution of $X_2|X_1$ as:

$$X_2|X_1 \sim F_2(\boldsymbol{\Theta}_1(X_1))$$

and the functional distributions of $\boldsymbol{\Theta}_1$ as:

$$\boldsymbol{\Theta}_1(X_1) = \boldsymbol{g_1}(\boldsymbol{\alpha}_{\mathbf{0}}^{(\mathbf{2})} + \boldsymbol{\alpha}_{\mathbf{1}}^{(\mathbf{2})} X_1)$$

We also have to set a Prior distribution for the parameters in the linear combination:

$$\boldsymbol{\alpha}^{(2)} \sim G_2$$

We proceed in the same way to define the other relevant distributions for the remaining

covariates. In particular, for $X_k|\boldsymbol{X}_{1,2,\ldots,k-1}$ we set the marginal Likelihood as:

$$X_k|\boldsymbol{X}_{1,2,\ldots,k-1} \sim F_k(\boldsymbol{\Theta}_{k-1}(\boldsymbol{X}_{1,2,\ldots,k-1}))$$

This formulation is general as the choice of F_k and of the functional form of $\boldsymbol{\Theta}_{k-1}(\boldsymbol{X}_{1,2,\ldots,k-1})$ simply depends on the type of covariate X_k. For continuous, binary, categorical or count variables, F_k is specified accordingly (e.g., Normal, Bernoulli, multinomial, Poisson), making the framework flexible for any variable type.

We then, define the functional distribution of $\boldsymbol{\Theta}_{k-1}(\boldsymbol{X}_{1,2,\ldots,k-1})$ as:

$$\boldsymbol{\Theta}_{k-1}(\boldsymbol{X}_{1,2,..,k-1}) = \boldsymbol{g_{k-1}}\left(\boldsymbol{\alpha_0^{(k)}} + \sum_{i=1}^{k-1} \boldsymbol{\alpha_i^{(k)}} X_i\right)$$

and the prior for $\boldsymbol{\alpha}^{(k)}$ as:

$$\boldsymbol{\alpha}^{(k)} \sim G_k$$

In the last lines, we define similarly the distribution of the outcome given the covariates and the exposure variable as:

$$Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$$

$$\boldsymbol{\Theta}(\boldsymbol{X}, Z) = \boldsymbol{g}\left(\boldsymbol{\alpha_0^{(K+1)}} + \sum_{i=1}^{K} \boldsymbol{\alpha_i^{(K+1)}} X_i + \boldsymbol{\beta} Z\right)$$

$$\boldsymbol{\alpha}^{(K+1)} \sim G_{K+1}$$

When we face a survival setting and we have a time-to-event outcome, not only Y has to be estimated but also the distribution of C, the censoring time distribution. We similarly define the model for this distribution as we did for the outcome variable, namely:

$$C|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}_C(\boldsymbol{X}, Z))$$

$$\boldsymbol{\Theta}_C(\boldsymbol{X}, Z) = \boldsymbol{g}_C \left(\boldsymbol{\alpha_0 C^{(K+1)}} + \sum_{i=1}^{K} \boldsymbol{\alpha_i C^{(K+1)}} X_i + \boldsymbol{\beta}_C Z \right)$$

$$\boldsymbol{\alpha}_C^{(K+1)} \sim G_{K+1}^C$$

Specifically, this is achieved by applying the model above to a new dataset in which the dichotomic variable indicating whether the event is censored or not has been switched in values (from zeros to ones and vice versa).

Once we have estimated the posterior distribution of the parameters of interest, we can sample new individuals from their predictive distribution with a realistic dependence structure that is coherent with the behavior retrieved from the RCT. Heuristically, this model takes the causal form of our data from previous studies as input and infers the conditional distributions of the different variables present in our model.

2.2.2.2 Unbalance observed confounders

In this section, we invert the *Assignment procedure* component in the TTE methodology. As we have previously mentioned, our aim is to introduce bias into the RCT so that the two populations become unbalanced, thereby mimicking realistic observational data. This can be done in many different ways, but to be able to test our methodologies, we must introduce what we mean by "unbalance"; how we measure it, and how we define a notion of near observations.

We know that a popular method for adjusting for confounders in the two populations is through matching methods (Stuart (2010)), which involve matching individuals with similar characteristics to obtain comparable populations. Matching methods, in other words, are strictly related to a notion of distance between individuals. The idea is to use these notions of distance to measure the overlap between the two populations.

Since the main strategy we will adopt to induce bias and unbalance populations will be to add or discard observations from data in different ways, defining a notion of distance is

very useful not only to measure the overlap between populations but also to choose which observations to add or discard to "separate" populations in a proper way. Note that this idea originates from "inverting" the related step in the Target trial emulation methodology. Specifically, since matching methods are used in TTE to adjust for confounding variables and reduce bias, inverting them allows us to introduce bias into data.

There are two main types of matching methods that we can use to define a measure of overlap and a notion of distance; the first is **propensity score matching** (PSM) (Austin (2011)). The PSM matches the control and treated populations with respect to each individual's propensity score, helping to construct two similar populations in terms of the probability of being treated. There are several methods to perform propensity score matching, including nearest neighbor, optimal full matching, and exact matching (Stuart (2010)).

The second is **Mahalanobis matching** (MM) (Rubin (2006)); it works by matching control and treated populations with respect to pairwise Mahalanobis distance. In this case, the standard MM sequentially matches each individual in the control group with the closest treated individual(s) (in terms of Mahalanobis distance, we introduced in §2.3.1).

Both notions of induced distances have pros and cons, but we decided to proceed with the Mahalanobis distance. This choice is based on the fact that in ITTE, we start from an RCT, where individuals have the same propensity score. Therefore, to discriminate between them and introduce bias (either increasing or reducing overlap), we must work on the individual's covariates' profile, rather than on the probability of being treated.

Since the Mahalanobis distance enables us to define close and distant observations, we will utilize this metric both to add or discard some of them to unbalance populations introducing bias inside data, and to measure population imbalance as anticipated in §1.2.1.1.

Other possibilities for introducing bias include separating populations based on their covariate values, for example, keeping a high value of a particular covariate in one pop-

ulation and a low value in the other. Another way is to explicitly define the conditional distribution of the exposure variable given the covariates.

2.2.2.3 Immortal time

As mentioned above, immortal time typically occurs when an individual enters the study and waits a certain amount of time before starting therapy and is subsequently labeled as treated. Immortal time refers to the precise span of time in which observation of follow-up has already begun, but treatment has not yet been assigned. This leads to the fact that for treated people, the event cannot occur during the entire specified period, resulting in a biased estimate of treatment efficacy. This usually involves an overestimation of therapy efficacy, posing a real threat of drawing incorrect conclusions from the data.

There are several ways to adjust for immortal time bias (Cox model with a time-dependent treatment variable, sequential Cox approach, etc.), but in this work, we focus mainly on *Prescription time distribution matching* (PTDM) (Karim et al. (2016)).

PTDM involves redefining time zero in both the ever-treated (patients who receive the treatment and may exhibit immortal time) and the never-treated (patients who have not received the treatment to date) populations. For each never-treated individual, we randomly select (with replacement) an ever-treated individual's immortal time and assign it as the new time zero for the selected never-treated individual. If the event of the selected never-treated individual occurs before the end of the matched immortal time, they are excluded from the study.

Again, if in the TTE procedure, we aim to eliminate time-zero bias, in ITTE, we invert the procedure to induce the relevant bias in the simulated data. To this end, we first describe how one can simulate the individual's immortal time for ever-treated patients. In standard scenarios, the conditional distribution of Immortal time given patients' covariates $It|\boldsymbol{X}$ can be informed by the literature:

$$It|\boldsymbol{X} \sim F_{It}(\boldsymbol{\Theta}_{It}(\boldsymbol{X}))$$

Where the functional form of the parameters can be assumed to be linear in covariates:

$$\boldsymbol{\Theta}_{It}(\boldsymbol{X}) = g_{It}\left(\alpha_{0It} + \sum_{i=1}^{K} \alpha_{iIt} X_i\right)$$

In case we possess some observational data collected in a similar environment (in the same hospital) as the data we want to simulate, immortal time distribution can be inferred by using a similar Bayesian model to the one in §2.2.2.1 with the choice of prior probabilities informed by experts' opinions. In this case, parameters $\boldsymbol{\alpha}_{It}$ from the previous model are to be considered as random, and we have to define a prior distribution on them:

$$\boldsymbol{\alpha}_{It} \sim G_{it}$$

Another important component contributing to time-zero-related bias in observational data is the timing of arrival in the study for the different individuals (Anisimov and Fedorov (2007), Carter (2004)). Simulating this variable is essential for different reasons: to obtain realistic non-experimental data, to predict the number of patients (in a sensitivity analysis context) belonging to a future study, and also to account for time trends (in studies with a long recruitment phase) related to changed recruiting criteria, changed therapies, and improvement of diagnostic methods. Since observational studies typically collect data from multiple locations, it is crucial to model the differences that may exist between these locations accurately in terms of the arrival process.

Related to individuals' arrival, we model this process as a Bayesian Poisson-Gamma hierarchical model with parameters informed by real data and expert opinions. Formally, for any fixed couple (t, i), and $i = 1, .., n$ with n number of different accrual spots, we model $N_t(\lambda_i)$, that is, the number of individuals arrived after time t in the accrual spot

k, as:

$$N_t(\lambda_k) \sim Pois(\lambda_k t)$$

$$\lambda_k \sim Gamma(\alpha, \beta) \tag{2.1}$$

$$\alpha \sim F_\alpha$$

$$\beta \sim F_\beta$$

with F_a, F_b priors for Gamma parameters.
In particular, since we want to make inferences on the arrival process of individuals, we want to model the waiting times process. In this case, the induced model is an exponential-gamma hierarchical model, formally:

$$T_i(\lambda_k) \sim exp(\lambda_k)$$

$$\lambda_k \sim Gamma(\alpha, \beta)$$

$$\alpha \sim F_\alpha$$

$$\beta \sim F_\beta$$

With priors informed by expert opinions. Note that, also in this case, we assume we can update priors using data from previously collected health registries. If we don't possess any data, we again set values for $\boldsymbol{\lambda}$, informed by the literature.

In the next section, we describe the main methods for performing Inverse target trial emulation, focusing on the inversion of the *Assignment procedures* and *Time zero* components. Before proceeding, we summarize the revisitation and inversion of the TTE process in the table below, highlighting the main components of Target Trial Emulation and Inverse Target Trial Emulation.

	TTE	**ITTE**
Eligibility Criteria Treatment Strategies Outcomes	Apply exclusion and inclusion criteria leaving out from D_{obs} who do not satisfy them; in the same way keep the data coherent with the treatment strategies of the trial we are targeting and with the primary and secondary outcomes we want to measure.	Retrieve from RCT the information on covariates' joint distribution, treatment strategies, and outcome's conditional distribution to generate new data coherent with the knowledge extracted from RCT. Finally, add some noise to data related to the three components.
Assignment Procedures	To obtain conditional exchangeability and balance between treated and untreated populations, different methods can be used, such as: *matching*, *inverse probability weighting*, *regression methods*, *g-methods*	We add bias into data, making the two arms unbalanced and reducing overlapping. We define different methods to build an unbalanced observational dataset with different degrees of overlap and different degrees of knowledge on the treatment allocation mechanism.

Time Zero	Different methods to solve the misalignment between time when eligibility criteria are met, treatment is assigned, and follow-up begins (prescription time-distribution matching, nested trials,..).	Simulate the arrival of patients, their grace period (immortal time), their survival outcome, and the censoring variable. Add bias by rescaling the survival outcome conditioning to the grace period or adding dependency between the outcome variable and the arrival time.

Table 2.1: Summary of the revisitation and inversion of the TTE process. Comparison between ITTE and TTE components

2.3 Methods

In this section, we delve into the details of the main methods we define to unbalance cofounders and introduce time-zero bias in ITTE. Note that every method uses the preliminary step of ITTE described in §2.2.2.1 to retrieve information from the RCT and sample new data (individuals) coherently. Each of the following methods requires different assumptions.

2.3.1 Unbalanced confounders

We describe in detail the different methods to unbalance cofounders in the two populations as we introduced in §2.2.2.2. Each of the following methods requires different assumptions

about how the treatment allocation mechanism works, i.e., which are the differences in the two populations.

Before we dive into the concrete implementations, we recall the notation introduced in §2.2.1. The general objective is to obtain a dataset D_{obs} with unbalanced populations, and, as we mentioned, the first common step is to simulate a larger RCT (in terms of sample size) $\tilde{D}$, retrieving the original dependence structure from D_{rct} using the model in §2.2.2.1.

2.3.1.1 Perfect knowledge method

First, we describe the method by which we assume to have perfect knowledge of how the distributions of the two populations differ and how they are generated.

As in the previous section, we start by learning the covariates and outcome distributions. Finally, since we assume to know exactly how the treatment allocation process works, i.e. which is the conditional distribution $Z|\boldsymbol{X}$, note that, in the presence of two groups ($Z = 0, 1$), this is equivalent to having perfect knowledge of the functional form of propensity score $e(\boldsymbol{X})$. In this scenario, we simulate an observational dataset through the learned distribution and the specified allocation process.

Perfect knowledge method Starting from an RCT we build an observational dataset D_{obs} in the following way:

- Learn the covariates distribution and the conditional distribution of the outcome Y given the treatment status and the covariates $(Z, \boldsymbol{X})$ using the model in §2.2.2.1.
- Assume a specified conditional distribution for $Z|\boldsymbol{X}$.
- Simulate an observational dataset D_{obs} sampling N individuals from the posterior distributions above and the specified distribution for $\boldsymbol{Z}|\boldsymbol{X}$.

Notice that this method does not allow for control for the level of overlap since perfect information on the allocation process leads to a specific level of this measure.

2.3.1.2 Covariate based methods

Then we introduced two different methods to separate the two populations based on the difference in covariates we assume they exhibit. For example, suppose that we assume a specific therapy is usually prescribed more often to older people than to younger ones in a given hospital; then we will replicate this characteristic in our data.

In this case, the necessary assumption for working with these models is that we know the differences between populations in terms of covariates; meaning that we have a prior, even vague, idea on how the covariates distribution differs in the two populations (e.g. one population being generally younger and the other older). We might retrieve these information by relying on the literature or previous observational data in our possession. These methods then work by "discarding" specific individuals I from $\tilde{D}$ to create an imbalance between the two populations. By "discard" we mean the removal of the selected individuals from the dataset.

Starting from $\tilde{D}$, the first method, called **Decoupling covariate** takes n specific covariates $\boldsymbol{X}^{(n)}$ and discards N individuals $\{I_1, .., I_N\}$ based on the values of $\boldsymbol{X}^{(n)}$ and obtains D_{obs}. Specifying how $\boldsymbol{X}^{(n)}$ characterize the differences in the two populations can be straightforward, e.g., higher values of $\boldsymbol{X}^{(n)}$ in the control group and lower values in the treated group, or a more complicated characterization.

The second method, **Decoupling Mahalanobis**, selects N individuals $\{I_1, .., I_N\}$ from $\tilde{D}$ based on the values of specific covariates $\boldsymbol{X}^{(n)}$ and for each of them it discards the closest individual belonging to the opposite population to obtain D_{obs}, where "closeness" is in terms of Mahalanobis distance.

The main difference between these two methods is related to the metric we use to select individuals to be discarded; in the first method, the notion of distance is only given by

the covariate value of the given individual, while in the second one, we consider not only the covariate value but also the Mahalanobis distance to the selected individual. Hence, in the first case, we only consider the covariates that separate the two populations. At the same time, in the second method, the other covariates also play a role in the process of discarding individuals and separating populations. Notice that in both methods, the greater the value of N (the number of discarded individuals), the lower the level of overlap becomes. Therefore, we can continue discarding individuals until the desired degree of overlap is reached. By *desired degree of overlap* here we refer to scenarios in which researchers want to reach a certain degree for different possible reasons: 1) they know which degree of overlap the data they want simulate have, so they want to reproduce them coherently; 2) they want to test methods to solve for confounding with different degrees of overlap to test their robustness in different scenario. The critical property we must ensure is that with these methods, one can control for the degree of overlap, that is, the degree of bias they wants data to reproduce. We now summarize the two methods:

Decoupling covariate

- Starting from D_{rct}, simulate $\tilde{D}$ using the model in §2.2.2.1.
- Select n covariates $\boldsymbol{X}^{(n)}$ and specify how they induce differences in control and treated groups.
- Select N individuals $\{I_1, .., I_N\}$ from $\tilde{D}$ belonging to the control group whose covariates values characterize the difference between populations.
- Remove the selected individuals from $\tilde{D}$ to obtain D_{obs}.

As we highlighted above, in this method, we assume we know which covariates produce the non-overlap between the two populations and have an approximate idea (usually related to extreme values) of how they distribute among the two groups. We do not

assume to know precisely which is the conditional distribution of the exposure variable given the observed confounders.

Decoupling Mahalanobis

- Starting from D_{rct}, simulate $\tilde{D}$ using the model in §2.2.2.1.
- Select n covariates $\boldsymbol{X}^{(n)}$ and specify how they induce differences in control and treated groups.
- Select N individuals $\{I_1, .., I_N\}$ from $\tilde{D}$ belonging to the control groups whose covariates values characterize the difference between populations.
- For each I_k, remove from the dataset the individual J_k belonging to the treated group with the minimum Mahalanobis distance from I_k and obtain D_{obs}.

Again, in this method, we assume to know which are the covariates that produce the non-overlapping between the two populations and to have an approximate idea (usually related to extreme values) of how they distribute among the two groups. We do not assume to know precisely which is the conditional exposure distribution given the observed confounders.

2.3.1.3 Randomized new datasest

Finally, we introduce a method that assumes no knowledge available about how the two populations are separated. Starting from an RCT D_{rct}, it works by first learning the covariates and outcome distributions $p(X_1, \ldots, X_k)$, $Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$ from model in §2.2.2.1, then the Mahalanobis matrix $\boldsymbol{M}$ (with pairwise distances) is constructed.

Second, two individuals $\{I, J\}$ with a prespecified Mahalanobis distance between covariates and belonging to different populations are selected. For instance, we can select the two farthest individuals or the couple corresponding to the 0.75 quantile of the dis-

tribution of Mahalanobis distances taken from $\boldsymbol{M}$.

Finally, obtain D_{obs} by sampling N new individuals $\{I_1, .., I_N\}$ for each of the two populations following $p(\boldsymbol{X})$ and $Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$ centered in I for new individuals of the same population of I and centered in J for new individuals of the same population of J with variance depending on the distance of I or J to the center of the covariates' distribution $\overline{\boldsymbol{X}}$. Again, if the two individuals we select are far apart from each other (e.g a quantile of 0.8), and the larger N (the number of sampled individuals) is, the lower the level of overlapping becomes. Therefore, we continue sampling individuals until the desired degree of overlapping is reached.

Note that this method is sensitive to the choice of I and J, but overall the more I and J are far from each other and the greater N the more populations are separated and bias is induced in the data.

We now summarise the just described method to obtain D_{obs} from D_{rct}:

Randomised new datasest

- Learn the distributions $p(\boldsymbol{X})$ and $Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$ using the model in §2.2.2.1.
- Construct the Mahalanobis matrix M with the pairwise distances between all the individuals.
- Select the two farthest (or with a prespecified distance) individuals I and J belonging to the two different populations;
- Sample the new individuals' population i.i.d. from a Bernoulli random variable.
- Sample new individuals' covariates from $p(\boldsymbol{X})$ centered in I or J.
- Sample new individuals' outcomes from $Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$ centered in I or J.
- Add the sampled individuals to D_{rct} and obtain D_{obs}.

2.3.2 Immortal time bias

Finally, we describe how to generate realistic observational data in terms of immortal time. We first learn the distributions of all relevant parameters from the model in §2.2.2.1 and from the model in §2.2.2.3.

Second, we sample a larger RCT $\tilde{D}$, where, initially, all individuals' outcomes are sampled, as they belong to the control population; i.e., we sample the outcomes from $Y|\boldsymbol{X}, Z = 0$ and censoring time $C|\boldsymbol{X}, Z$. This is because we want to reproduce outcomes during immortal time, when follow-up has already started but people have not yet received

treatment; therefore, $Z = 0$.

Third, for each individual i of $\tilde{D}$, we sample their immortal time from $It|\boldsymbol{X}_i$, that is, the precise span of time in which follow-up observation has already started, but treatment has not yet been assigned.

Fourth, if i belongs to treated population and their immortal time is lower than $min(Y_i, C_i)$ we set $Y_i = It_i + min(y_i, c_i)$ where y_i is now a sample from $Y|Z = 1, \boldsymbol{X}_i$ and c_i is a sample from $C|Z = 1, \boldsymbol{X}_i$; representing the fact that after the immortal time, individual i actually started the therapy of interest (labeled as treated) and their outcome and censoring must now be sampled from $Y|Z = 1, \boldsymbol{X}_i$ and $C|Z = 1, \boldsymbol{X}_i$ respectively.

Finally, if the individual i belongs to the treated population and their immortal time is greater than $min(Y_i, C_i)$, then we set $Z_i = 0$, indicating that this individual is considered untreated as either censoring or time to event occurs before they could have received the treatment. This final step is a key element that leads to bias, as it is clear from here that an individual to be considered treated must at least survive until their immortal time (leading to an overestimation of efficacy).

We now summarize the just-described process.

Immortal time bias

- Starting from D_{rct}, simulate $\tilde{D}$ using the model in §2.2.2.1 with each individual's outcome sampled from $Y|\boldsymbol{X}, Z = 0$ and censoring time from $C|\boldsymbol{X}, Z = 0$.
- For each individual I of $\tilde{D}$ sample immortal time It_I from §2.2.2.3.
- If $Z_I = 1$ and $It_I < min(Y_I, C_I)$ set $Y_I = It_I + min(y_I, c_I)$ with novel y_I and c_I sampled this time from $Y|Z = 1, \boldsymbol{X}_I$ and $C|Z = 1, \boldsymbol{X}_I$.
- If $Z_I = 1$ and $It_I > min(Y_I, C_I)$ set $Z_I = 0$.

2.4 Test ITTE induced bias and TTE robustness

In the previous sections, we introduced Inverse target trial emulation and defined various methods for performing ITTE in realistic observational settings. This novel methodology aims to simulate realistic data from experimental data and can be useful in various scenarios.

We aim to test two aspects of the methodology described above: first, its capacity to produce a dataset with a previously defined level of bias (in terms of unbalanced population and time zero bias), recognizing that this ability is essential for sensitivity analysis and research prioritization in various realistic settings (e.g. VoI methods). Second, the ability of ITTE to assess the robustness of TTE methods in various scenarios. Starting from an RCT, this second test can be performed by sequentially applying ITTE and TTE, and comparing some relevant estimates (such as the average treatment effect (ATE) and mean survival difference (MSD)) from the initial RCT with those from the final emulated ones. For the complete list of relevant parameters and distributions we will assume in the following tests, see §7.1.

2.4.1 Testing ITTE in generating bias: unbalanced populations

We begin by testing the capacity of ITTE to recreate realistic biases within the data. The first test is related to inducing imbalance in populations; that is, we aim to test the capacity of ITTE in unbalancing the treatment and control groups.

From the previous section, it is clear that population can be unbalanced by either sampling, in case of *Randomised new dataset* method §2.3.1.3, or discarding, in case of *Covariate based methods* §2.3.1.2), an increasing number of individuals. The methods introduced in the previous section are defined in a way that suggests that the more individuals we sample or discard, the more the amount of bias in the data increases. In this context, we will measure bias in two different ways: 1) as population Mahalanobis distance

between groups, 2) as the difference in relevant efficacy estimates (Average treatment effect, Mean survival difference, etc.) between the original RCT D_{rct} and the observational one D_{obs} obtained through ITTE.

Specifically, we will generate multiple observational datasets by applying the different ITTE methods from the previous section multiple times to an initial RCT dataset. Each time we apply ITTE methods, we sample/discard an increasing number of individuals and test whether the population overlap and the difference between the original RCT and the generated data in treatment efficacies vary accordingly.

We apply this test to the German Breast Cancer Study Group (GBSG2) dataset (Schumacher et al. (1994)), a study examining the effects of hormone treatment on recurrence-free survival time. The data is composed of 686 observations: from various individual characteristics (*age, menopausal status, tumor size, tumor grade, number of positive nodes, progesterone receptor, estrogen receptor*), to exposure variable, recurrence-free survival time as measured outcome, and the related censoring indicator.

For these reasons, to test our methodology in the context of unbalanced confounders, we will proceed in the following way:

Generating bias - Unbalance Confounders

1. Compute mean survival difference (MSD) (or other quantities of interest related to treatment efficacy) in the original RCT D_{rct}.
2. Generate observational data $\{D_{obs}^{(N_1)}, .., D_{obs}^{(N_n)}\}$ performing one of the ITTE methods multiple times, every time with a different number of individuals (N_i) removed or added.
3. Compute the overlapping degree in each $D_{obs}^{(N_i)}$, as Mahalanobis population distance.
4. Compute multiple MSDs in $\{D_{obs}^{(N_1)}, .., D_{obs}^{(N_n)}\}$ and compare these results with the original MSD depending on the overlapping degrees of $\{D_{obs}^{(N_1)}, .., D_{obs}^{(N_n)}\}$.
5. Repeat the above algorithm m times and take the mean of the relevant estimates.

In the next section, we perform this test by considering on point 2 of the above box, *Decoupling covariate* and *Randomised new dataset* as the ITTE methods to perform. We generate $\tilde{D}$ of 1000 individuals and progressively discard/add $N = 30, 60, 90, 120, 150$ individuals to obtain $\{D_{obs}^{(30)}, .., D_{obs}^{(150)}\}$. Moreover, to ensure the validity of the test, we will repeat this process $m = 10$ times in the next section. The results are shown in the next section.

2.4.2 Testing ITTE in generating bias: immortal time

The second test related to testing the ability of ITTE in generating bias involves the incorrect definition of time zero; that is, we aim to assess the capacity of ITTE in inducing immortal time bias within the data.

In §2.3.2 we defined how to simulate observational data with immortal time bias. We

can measure the amount of immortal time bias induced in two ways; first, we count the misclassified individuals, who are those assigned to the control group, even if they were supposed to receive treatment, because they were not in time to receive it. Second, as in the previous test, we measure the difference in relevant efficacy estimates (Average treatment effect, Mean survival difference, etc.) between the original RCT D_{rct} and the observational one D_{obs} obtained through ITTE.

If in the case of inducing population imbalance, we could sample/discard more individuals to induce more bias; in this case, more bias is induced by sampling higher immortal time values with higher probability. In particular, in the next section, we will sample immortal time in five different ways, i.e., from five different Gamma distributions with the same rate but increasing shape.

To test the introduced methodology in terms of immortal time bias, we will proceed as follows:

Generating bias - Immortal time

1. Estimate the MSD in the two populations (or other quantities of interest related to treatment efficacy) in the original RCT D_{rct}.
2. Simulate a bigger RCT $\tilde{D}$ using the model in §2.2.2.1.
3. Induce immortal time as outlined in §2.3.2, sampling immortal time from n different distributions (or the same distribution with n different parameters) obtaining $\{D_{obs}^{(1)}, .., D_{obs}^{(n)}\}$.
4. Estimate MSDs of $\{D_{obs}^{(1)}, .., D_{obs}^{(n)}\}$ and compare these results with the different distributions (or the different distribution's parameters) chosen for modeling immortal time.
5. Repeat the above algorithm m times and take the mean of the relevant estimates.

To test the ITTE methodology in terms of induced bias, we will model immortal time using $n = 5$ Gamma random variables with a constant rate $\lambda = 0.5$ and increasing shape $\boldsymbol{\alpha} = (20, 40, 60, 80, 100)$. Moreover, to ensure the validity of the test, we will repeat this process $m = 10$ times in the next section.

Before proceeding to describe the results, we outline another test we conducted to assess the importance of Inverse target trial emulation.

2.4.3 Testing the robustness of TTE

As we mentioned at the beginning of this section, simulating observational data starting from RCT can be useful for different reasons. One of these involves testing the robustness of different methods used for TTE in adjusting for various biases. In this context, we focus on testing the ability of different methods to solve the bias linked to Time zero

definition and Assignment procedure as we introduced in §1.2.1.3.

The idea behind these tests is, starting from RCT data D_{rct} and some associated relevant measures (average treatment effect, mean survival difference,..) $\theta(D_{rct})$, apply the above ITTE methods to generate different D_{obs} in different realistic settings and then apply the TTE methods we want to test targeting D_{rct} and obtain $\theta(\tilde{D}_{rct})$. Finally, we can compare $\theta(D_{rct})$ with $\theta(\tilde{D}_{rct})$ to evaluate the TTE methodology in retrieving the original RCT data measures.

In this context, there are more possible applications of ITTE. We can either compare different TTE methods to evaluate their robustness, or test the robustness of one specific method in the presence of varying degrees of bias, or we can do both simultaneously. As we mentioned, the following sections define two different tests. The first test aims to evaluate the robustness of PTDM in adjusting for immortal time bias with varying degrees of bias. The second test compares different methods to adjust for confounding and balance populations.

2.4.3.1 Testing TTE: Time zero

First, ITTE can be useful for testing the robustness of the different TTE methods in adjusting immortal time bias. Since the definition of Time zero is a crucial step in a proper TTE process, testing its robustness in different contexts becomes necessary.

To perform this test, we start by estimating the MSD in the initial RCT. Then, as above, we simulate a larger RCT using the model in Section §2.2.2.1. The next step is to induce immortal time bias, as outlined in §2.3.2, by sampling immortal time from n different distributions that induce varying degrees of bias, thereby obtaining n distinct observational datasets. Finally, we apply PTDM (or another method we want to test) targeting the initial RCT, and estimate MSDs of the targeted trials. We then compare the original MSDs and the final ones. To ensure the validity of the test, we will repeat this process $m = 1000$ times in the next section.

In the next section, we apply this test again to the GBSG2 dataset and present the results.

Testing TTE: Time zero

1. Estimate the MSD in the two populations (or other quantities of interest related to treatment efficacy) in the original RCT D_{rct}.
2. Simulate a bigger RCT $\tilde{D}$ using the model in §2.2.2.1;
3. induce immortal time as outlined in §2.3.2, sampling immortal time from n different distributions (or the same distribution with n different parameters) obtaining $\{D_{obs}^{(1)}, .., D_{obs}^{(n)}\}$.
4. Estimate MSDs of $\{D_{obs}^{(1)}, .., D_{obs}^{(n)}\}$.
5. Use one of the TTE methods, such as PTDM, to compute the MSD of the emulations of the starting RCT $\{\tilde{D}_{rct}^{(1)}, .., \tilde{D}_{rct}^{(n)}\}$.
6. Compare all the different estimates.
7. Repeat the above algorithm m times and take the mean of the relevant estimates.

2.4.3.2 Testing TTE: Assignment procedure

If we want to test another crucial step in TTE, i.e., the assignment procedure step, we can proceed similarly.

In particular, starting from an RCT, we can compute an unbiased estimate of the average treatment effect or other quantities of interest. Then we sample different observational data, introducing bias in different ways by exploiting all ITTE methods in §2.3, and observe how the amount of bias introduced (in unbalanced populations) translates

into a change in ATE. At this point, we can apply different TTE methods to adjust for confounding, targeting the initial trial, and computing ATE. Finally, by comparing the ATE of the emulated trial with that of the original one, we can evaluate the robustness of the tested methods in adjusting for different types of confounding. To ensure the validity of the test, we will repeat this process $m = 1000$ times in the next section.

We perform the proposed test on a modified version of the *Help Dataset* Samet et al. (2003): a clinical trial for adult inpatients recruited from a detoxification unit. Patients with no primary care physicians were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the aim of linking them to primary medical care. The data comprise 453 observations, ranging from various individual characteristics: *gender, days to first drink since leaving detox, SF-36 Physical and Mental components, depression score, number of days to first substance since leaving detox, risky sexual behavior, risky drug-related behaviors*), as outcome *days to linkage to primary care*, and the exposure variable.

Since we only want to test our methodology related to the induced imbalance in observed confounders in this setting, we simplify the approach to this dataset by excluding the survival component of the data above, considering the outcome as a general measure of therapy effectiveness.

In particular, to test our methods and to use them to verify the robustness of classical TTE methods with respect to adjusting for unbalanced confounders, we proceed in the following way:

Testing TTE: Assignment procedure

1. Compute ATE (or other quantities of interest related to treatment efficacy) in the original RCT D_{rct}.
2. Run the ITTE methods n times and obtain different observational datasets $\{D_{obs}^{(1)}, .., D_{obs}^{(n)}\}$ for each method.
3. Compute multiple ATEs in the observational datasets without any kind of transformations and make a comparison both with the original one and one to each other (in terms of mean ATE and absolute error).
4. Use Mahalanobis matching, Inverse probability weighting, regression based methods and augmented inverse probability weighting into a TTE process targeting the initial RCT and compute the ATE in the emulations of the starting RCT $\{\tilde{D}_{rct}^{(1)}, .., \tilde{D}_{rct}^{(n)}\}$.
5. compare all the different estimates.
6. Repeat the above algorithm m times and take the mean of the relevant estimates.

Note that in this test, we focus our attention on four specific causal inference methods among the most common ones, even if this procedure can be used to potentially test the robustness of any causal inference method. Again, for the complete list of relevant parameters and distributions assumed in the applications, see §7.1.

2.5 Results

We now present the results of the tests described above, using the same terminology as we refer to the specific test being performed in each subsection. First, we test our

methodology in the context of the ability of the ITTE methodologies to generate bias. Second, we test the robustness of different TTE methods exploiting ITTE.

2.5.1 Generating bias - Unbalancing confounders

We test the level of induced bias in relation to the level of induced overlap between the two populations. Among all the previously introduced methods, the ones that allow control for the level of overlapping are: Decoupling Mahalanobis, Decoupling covariate and Randomised new dataset. We focus on the last two methods and present here two different plots representing the bias (measured in difference in mean survival difference with respect to the starting MSD) as a function of the level of overlap.

In the following plots, we test our ability to control for the level of non-overlapping and for the difference in generated therapy efficacy as described in §2.4.1. First, we run the ITTE method in §2.3.1.3, i.e. *Randomised new dataset*.

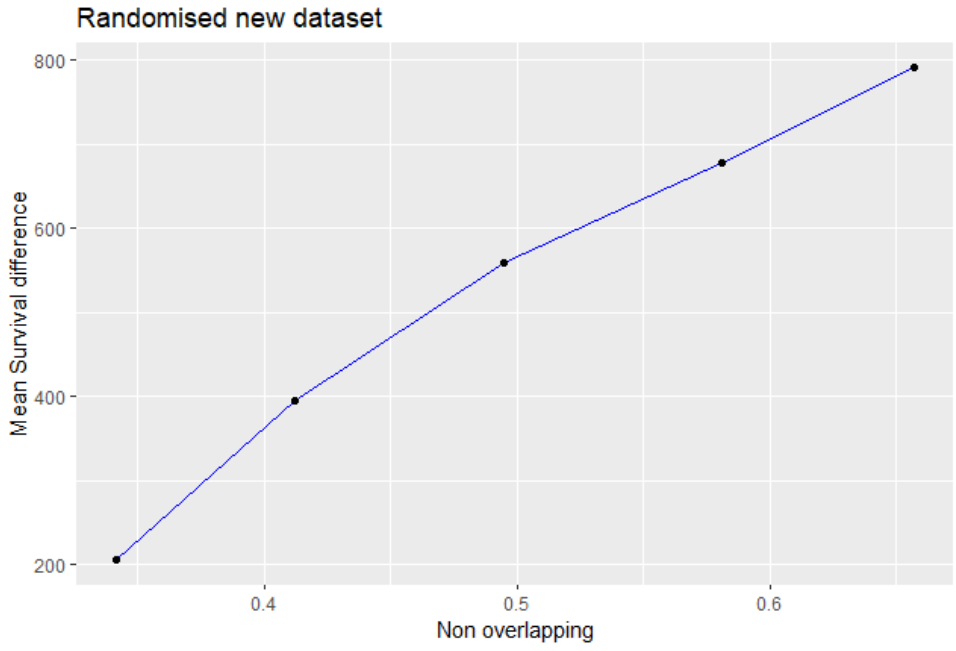


Figure 2.1: Induced level of non-overlapping (populations Mahalanobis distance) applying *Randomised new dataset* method to obtain $\{D_{obs}^{(30)}, .., D_{obs}^{(150)}\}$ against difference in MSD

Second, we run the ITTE method in 2.3.1.2, i.e. *Decoupling Covariate*.

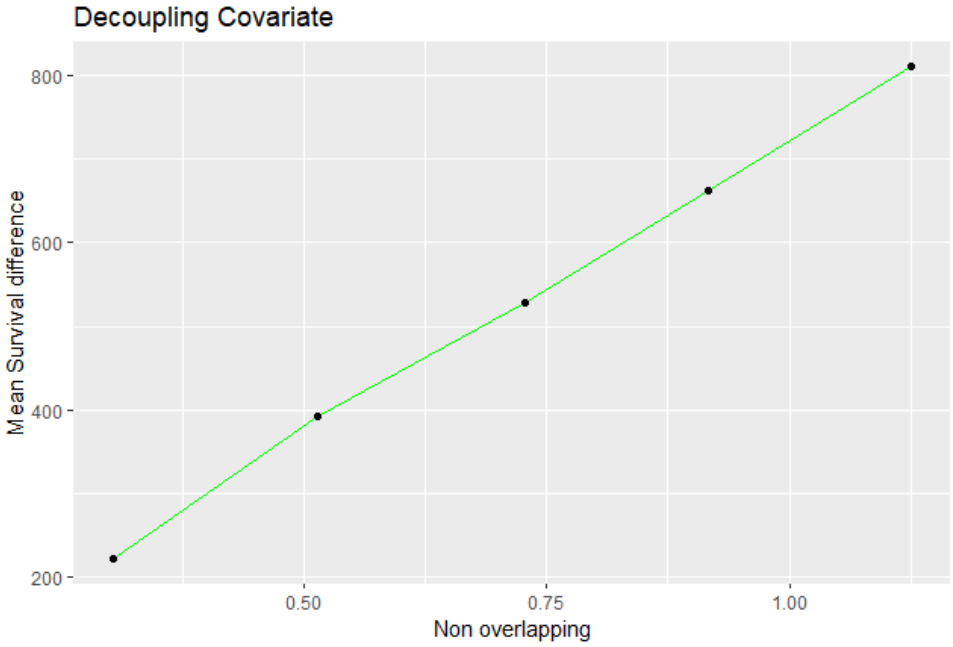


Figure 2.2: Induced level of non-overlapping (populations Mahalanobis distance) applying *Decoupling covariate method* method to obtain $\{D_{obs}^{(30)}, .., D_{obs}^{(150)}\}$ against difference in MSD

As one can see, both methods show that we can control for the level of bias induced in the generated data (through the number of individuals added or discarded) under different degrees of knowledge and various methodologies.

2.5.2 Generating bias - Immortal time

We now move on to testing ITTE with respect to Time zero definition, to understand if we can control for the bias we induce inside the data. The results are presented in Table 2.2 in terms of induced bias (measured in difference mean survival time) in the dataset, noting that **the original estimated mean survival difference is 874 days**.

As we can observe, in this case also, we are able to control for the level of induced bias (both in terms of misclassified individuals and differences in generated therapy efficacy) through the parameters of the distribution that simulate Immortal time. As we expected, the higher the immortal time values, the more individuals are misclassified.

	original RCT	shape = 20 rate = 1	shape = 80 rate = 2	shape = 180 rate = 3	shape = 320 rate = 4	shape = 500 rate = 5
mean survival difference	874	876.40	981.67	1085.80	1181.63	1276.31
misclassified individuals	0	10	20	30	39	49

Table 2.2: Induced immortal time bias is measured both as the number of missclassified individuals and MSD. Immortal time is generated with five different Gamma random variables with a constant rates of (1,2,3,4,5) and shapes of (20,40,60,80,100).

2.5.3 Testing TTE: Time zero

We now test for the robustness of TTE methods, focusing our attention on PTDM to adjust for immortal time bias in the same context as above. Table 2.3 shows the difference in mean survival time in simulated data between populations before and after applying PTDM.

	original RCT	shape = 20 rate = 1	shape = 80 rate = 2	shape = 180 rate = 3	shape = 320 rate = 4	shape = 500 rate = 5
mean survival difference	874	876.40	981.67	1085.80	1181.63	1276.31
mean survival difference PTDM		775.62	776.55	782.08	785.46	788.70
misclassified individuals	0	10	20	30	39	49

Table 2.3: Testing the robustness of PTDM in properly accounting for immortal time bias. Induced immortal time bias is measured both as the number of missclassified individuals and MSD. Immortal time is generated with five different Gamma random variables with a constant rates of (1,2,3,4,5) and shapes of (20,40,60,80,100).

As we can see, PTDM seems to underestimate treatment efficacies for the analyzed degrees of bias, that is, in the presence of relatively low values of misclassified individuals (less than 10% of missclassified individuals). For this reason, using sequential Cox approach might be better in these types of scenarios (Karim et al. (2016)).

This example also highlights the importance of understanding how to control the level of overlap (and therefore bias) to reproduce different situations and analyze the risks associated with performing TTE in various contexts.

2.5.4 Testing TTE: Assignment procedure

The TTE methods we decided to test for robustness in terms of adjusting for confounding are the most commonly used in this context: Full Mahalanobis matching, Inverse probability weighting (IPW), regression method, and Augmented inverse probability weighting (AIPW).

The following tables present the results in terms of average treatment effects, with variance in brackets.

	Original RCT	Rand new dataset	Dec Mahalanobis	Perfect knowledge	Dec covariate
Original Data	125.19 (1.64)	154.26 (1.50)	124.05 (0.98)	76.28 (5.52)	88.72 (15.86)
Mahalanobis Matching		149.56 (8.01)	126.11 (0.65)	124.51 (1.61)	125.70 (1.13)
Inverse Probability Weighting		115.65 (1.80)	140.13 (14.25)	106.12 (8.25)	95.80 (32.20)
Regression Based Adjustment		130.02 (0.50)	125.33 (0.30)	124.10 (1.36)	125.84 (0.63)
Augmented Inverse Probability Weighting		127.76 (0.66)	125.55 (0.22)	124.00 (1.50)	125.70 (0.88)

Table 2.4: ATE estimates given by Mahalanobis matching, IPW, regression methods, and AIPW with bias induced by methods described in §2.3, standard deviation in brackets.

We also report the mean absolute errors for each method:

	Randomised new dataset	Decoupling Mahalanobis	Perfect knowledge	Decoupling covariate
Original Data	29.08	1.17	48.90	36.47
Mahalanobis matching	24.38	0.94	1.38	0.85
Inverse Probability Weighting	9.54	14.94	19.16	29.52
Regression Based Adjustment	4.83	0.29	1.41	0.68
Augmented Inverse Probability Weighting	2.57	0.39	1.54	0.71

Table 2.5: Mean absolute errors of relevant ATE estimates above.

As we can observe, AIPW and regression methods are the most robust in terms of both estimation and mean absolute errors. This is expected as in those methods we specify the same outcome model that we specify in §2.2.2.1. Moreover, most of the ITTE methods

induce bias through covariates and their distance; for this reason, methods as standard IPW, when the induced bias is high, have to account for a considerable level of variability, which translates into huge weights in IPW procedure and way poorer estimates. The full Mahalanobis results are similar to those of AIPW and regression-based method, except for the Randomized new dataset method, which inherently contains the largest amount of variability and bias. Overall AIPW performs slightly better than the other methods due to its double robustness property that we will test in the next section.

2.5.4.1 Double robustness of AIPW

We have mentioned that methods that specify the outcome model work better if we use the same model to generate $\tilde{D}$, as the one defined in §2.2.2.1. Despite this, AIPW is supposed to be doubly robust (Kurz (2022)), meaning that one has to specify either the outcome or the propensity score model correctly. We will test this property using ITTE by assuming a different outcome model for both the regression method and the AIPW. In particular, we misspecify the outcome model; if in the previous application we assumed a normal distribution to model the outcome, this time we repeat the previous experiments by assuming a gamma distribution for both the regression-based method and the AIPW.

The following table presents the analysis results, including both ATE estimates (with their corresponding 95% confidence intervals) and absolute mean errors.

	Original RCT	Rand new dataset	Dec Mahalanobis	Observational Bayes	Dec covariate
Original Data	125.19	154.26	124.05	76.28	88.72
Regression Based Adjustment		131.70	124.13	76.67	82.37
		(129.81 ; 133.36)	(123.01 ; 124.99)	(66.44 ; 86.30)	(55.52 109.80)
Augmented Inverse Probability Weighting		125.14	126.33	118.63	108.76
		(123.40 ; 126.83)	(125.62 ; 128.16)	(111.10 ; 126.32)	(77.90 ; 125.42)

Table 2.6: ATE estimates and 95% confidence intervals under outcome model misspecification

	Randomised new dataset	Decoupling Mahalanobis	Observational Bayes	Decoupling covariate
Regression Based Adjustment	6.51	1.05	48.51	42.82
Augmented Inverse Probability Weighting	0.70	1.15	6.74	16.47

Table 2.7: Absolute mean errors under outcome model misspecification

As we can observe, the doubly robust property of AIPW seems to hold except for the cases, already highlighted above, where the IPW has the poorest estimation performance (*Observational Bayes* and *Decoupling Covariate*). With respect to *Decoupling Mahalanobis* method, even if the IPW did not exhibit a great performance, the regression based adjustment method performs decently and so does the AIPW. In the most problematic scenarios (*Observational Bayes* and *Decoupling Covariate*), AIPW included the true value in the 95% confidence intervals but exhibits a huge variance and a poor estimate. This behavior is coherent with the theory as the AIPW estimates is unbiased only if at least one between IPW and regression based estimators is.

In a similar way, we can test the other side of the double robustness property, specifically, when the propensity score model is misspecified. To do this, we estimate the propensity score using a logistic regression model that intentionally omits two covariates known to induce confounding. These omitted variables are included in our outcome model but excluded from the propensity score model, thereby introducing misspecification and allowing us to evaluate whether the method retains consistency under this misspecification. The following are the results presented in the same way as above.

	Original RCT	Rand new dataset	Dec Mahalanobis	Observational Bayes	Dec covariate
Original Data	125.19	154.26	124.05	76.28	88.72
Inverse Probability Weighting		116.12	123.12	76.29	83.48
		(112.21 ; 119.80)	(122.09 ; 123.72)	(67.23 ; 85.53)	(56.45 ; 110.17)
Augmented Inverse Probability Weighting		127.77	125.48	124.09	125.80
		(126.55 ; 129.12)	(124.80 ; 125.82)	(121.31 ; 126.64)	(125.00 ; 127.33)

Table 2.8: ATE estimates and 95% confidence intervals under propensity score model misspecification

	Randomised new dataset	Decoupling Mahalanobis	Observational Bayes	Decoupling covariate
Inverse Probability Weighting	9.07	2.07	48.89	41.71
Augmented Inverse Probability Weighting	2.58	0.36	1.42	0.65

Table 2.9: Absolute mean errors under propensity score model misspecification

As we can see AIPW is much more robust under propensity score model misspecification, estimating in a correct way the true ATE in the 95% confidence intervals (with a tiny exception for the first model); while IPW completely fails in estimating the correct efficacy.

2.6 Conclusions

The purpose of this work has been to introduce and illustrate a methodology that enables the construction of a realistic observational dataset from a randomised clinical trial. To this aim, we introduced Inverse Target trial emulation, inverting all the relevant steps of Target trial emulation.

This methodology can be useful for different reasons, such as testing different TTE methods, performing sensitivity analyses across a range of scenarios, and understanding which analytical approaches are most appropriate for observational data subject to different types of bias. Beyond these primary purposes, that we already explored throughout

this chapter, it is worth considering additional applications. For example, ITTE methodology could be used to model the joint distribution of patients covariates required for microsimulation models. In settings where individual-level data are incomplete or unavailable, the ability to generate realistic multivariate covariate structures may substantially support the development of synthetic patient populations and improve the accuracy of subsequent simulations; this application needs further investigation in concrete scenarios.

Starting from defining a Bayesian methodology to infer the dependence relations between variables in the RCT, we defined concrete methods to perform ITTE under different degrees of prior knowledge of the bias contained in the observational data. We focused on cases of imbalance in treated and controlled populations, as well as immortal time bias. It would be interesting to define other types of methods that can model bias different from the ones proposed in this work, such as additional time-related biases, information or measurement bias, interval censoring, or bias linked to disease severity.

Due to the specific purpose of this work, we did not test all possible types of TTE methods, but only a subset among the most common ones, we introduced in §1.2.1.3. A specific study on testing the vast variety of TTE methods in the presence of different types of bias using ITTE would be particularly interesting. Note also that the primary aim of this work was to introduce the ITTE methodology and apply it in various contexts. A more accurate and detailed analysis is needed to examine how different methods perform in the face of various types and degrees of bias, as well as to investigate ITTE usage in concrete studies.

In the introduction, we mentioned that the novel methodology partially overcomes the issues related to the two primary data-generating mechanisms in this context, which are resampling with replacement and explicit modeling. We said 'partially' because, even if we are able to generate observational data without relying on the same dataset and without explicitly modeling the process in which bias is induced, in the Bayesian model in §2.2.2.1 we are still assuming a specific parametric distribution for the different variables

of the model. For this reason, a crucial direction to investigate is related to inferring the relevant dependencies relying on nonparametric models, machine learning methods, or nonparametric copulas. In this way, we would further prove the potential of ITTE in incorporating different methods that work under different hypotheses. For an interesting discussion on a related topic and some possible interesting methods to be included in the ITTE methodology, in particular in the step related to extrapolation of dependence structure from the initial trial, see (Dorie et al. (2019)).

References

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

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Austin, P. C. and Schuster, T. (2016). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical methods in medical research*, 25(5):2214–2237.

Bottigliengo, D., Baldi, I., Lanera, C., Lorenzoni, G., Bejko, J., Bottio, T., Tarzia, V., Carrozzini, M., Gerosa, G., Berchialla, P., et al. (2021). Oversampling and replacement strategies in propensity score matching: a critical review focused on small sample size in clinical settings. *BMC medical research methodology*, 21:1–16.

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

Concato, J., Shah, N., and Horwitz, R. I. (2017). Randomized, controlled trials, observational studies, and the hierarchy of research designs. In *Research Ethics*, pages 207–212. Routledge.

Desai, R. J., Wyss, R., Abdia, Y., Toh, S., Johnson, M., Lee, H., Karami, S., Major, J. M., Nguyen, M., Wang, S. V., et al. (2019). Evaluating the use of bootstrapping in cohort studies conducted with 1: 1 propensity score matching—a plasmode simulation study. *Pharmacoepidemiology and Drug Safety*, 28(6):879–886.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition.

Fu, E. L., Evans, M., Clase, C. M., Tomlinson, L. A., van Diepen, M., Dekker, F. W., and Carrero, J. J. (2021). Stopping renin-angiotensin system inhibitors in patients with advanced ckd and risk of adverse outcomes: a nationwide study. *Journal of the American Society of Nephrology*, 32(2):424–435.

García-Albéniz, X., Hsu, J., and Hernán, M. A. (2017). The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *European journal of epidemiology*, 32:495–500.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764.

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., and Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology*, 79:70–75.

Hilton Boon, M., Burns, J., Craig, P., Griebler, U., Heise, T. L., Vittal Katikireddi, S., Rehfuess, E., Shepperd, S., Thomson, H., and Bero, L. (2022). Value and challenges of using observational studies in systematic reviews of public health interventions.

Karim, M. E., Gustafson, P., Petkau, J., Tremlett, H., Benefits, L.-T., of Beta-Interferon for Multiple Sclerosis (BeAMS) Study Group, A. E., Ehsanul Karim, M., Gustafson, P., Petkau, J., Tremlett, H., Shirani, A., et al. (2016). Comparison of sta-

tistical approaches for dealing with immortal time bias in drug effectiveness studies. *American journal of epidemiology*, 184(4):325–335.

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167.

Lévesque, L. E., Hanley, J. A., Kezouh, A., and Suissa, S. (2010). Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *Bmj*, 340.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Samet, J. H., Larson, M. J., Horton, N. J., Doyle, K., Winter, M., and Saitz, R. (2003). Linking alcohol-and drug-dependent adults to primary medical care: a randomized controlled trial of a multi-disciplinary health intervention in a detoxification unit. *Addiction*, 98(4):509–516.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., and Rauschecker, H. (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Suissa, S. (2008). Immortal time bias in pharmacoepidemiology. *American journal of epidemiology*, 167(4):492–499.

Chapter 3

Calculating the Expected Value of Sample Information accounting for missing data

3.1 Introduction

The Expected Value of Sample Information (EVSI) (Ades et al. (2004), Baio (2013)) quantifies the economic value of collecting additional evidence (data) to reduce uncertainty in an underlying health-economic model. In general, the collection of new data results in a decrease in the uncertainty surrounding the parameters that characterize the economic model, ultimately leading to an update of knowledge on which treatment is optimal in terms of cost-effectiveness. The EVSI quantifies how valuable this process is in terms of the economic value of the additional information.

For the reasons mentioned above, EVSI has been demonstrated to be a powerful tool for determining whether collecting new data is worthwhile in many realistic applications. It allows us not to waste resources on treatments that appear to be effective only due to a lack of information, or to invest money in research for treatments that have already been evaluated with sufficient precision. At the same time, it is a valid instrument to improve the design of a study by understanding: 1) which data to collect to reduce the uncertainty in the most important parameters of the economic model (research prioritization) and 2) how much additional data must be collected to reach a certain level of uncertainty and maximize the additional value of information given by the data while minimizing the correlated costs.

As we highlighted in the introduction , EVSI has been applied only to fairly simple data collection exercises (Heath et al. (2022)). In most cases, it has been used to understand the value of randomized clinical trial (RCT) data, meaning that it measured the value of performing additional RCTs to reduce uncertainty.

RCTs are the gold standard in health technology assessment (HTA) because they minimize various forms of bias, including confounding, selection, and information bias. One of the limitations of computing EVSI with "idealized" RCTs is that the process does not account for the fact that RCTs, in realistic applications, often exhibit missing data.

Understanding how to perform EVSI with missing data can be beneficial, as it enables us to utilize this powerful instrument in a more realistic context. Moreover, like in the standard procedure, understanding how to apply EVSI in this context would be helpful to design experiments when we are planning to collect data that suffer from missingness (research prioritization).

EVSI computation with data affected by missingness has been performed in (Heath et al. (2022)) only in the context of missing completely at random (MCAR), which is highly unrealistic in real-world studies. In this work, we will develop a novel methodology to compute EVSI in a more realistic context with missing data.

To compute EVSI in this context, we start by generating complete RCTs as individual-level data and imposing missingness in different forms and under different assumptions. Then, we proceed in the same way as we usually do when dealing with missing data, that is, we analyze data using various techniques to account for missingness (Little and Rubin (2019), Rubin (2018), Seaman and White (2013)). Finally, we compute EVSI on the "adjusted" data. In this work, we will perform multiple imputations to recover complete RCTs that we can use to compute EVSI.

When we obtain EVSI in this novel context, we will compare the results obtained by performing EVSI on the original RCT and on the data we obtain through multiple imputations. This will help us understand the possible drawbacks of using EVSI with these kinds of data and how to address them.

The chapter proceeds as follows: first, we recall the definition of EVSI in Section §1.3.1 and describe how to generate data with missingness under different assumptions, as well as how missingness can be addressed using multiple imputations. Second, we introduce a general setting with an underlying health-economics model and a Bayesian data-generating model to compute EVSI with missing data.

Third, we apply this methodology in two different health-economic settings. In the first, we will assume a conjugate Normal Normal model for the data-generating process.

In contrast, the second scenario refers to a realistic setting where the underlying economic model is the Markov model introduced in §1.1.2.

Finally, we represent and discuss the EVSI of the initial RCT and the EVSI of the final imputed data in different scenarios, and propose some possible solutions to effectively utilize this methodology and more accurately apply EVSI methods in a realistic context.

3.2 Standard EVSI and missing data simulation

We recall the decision-analytic framework introduced in §1.1.1. Consider an underlying health economic model governed by parameters $\boldsymbol{\theta}$, which we assume to have a joint distribution p that represents the uncertainty associated with the set of parameters.

In this context, we assume that we need to compare a set of D different decisions, $d = 1, \ldots, D$, each of which is associated with a specific net benefit $\mathrm{NB}_d(\boldsymbol{\theta})$. Recall that the definition of the EVSI, introduced in §1.3.1, is

$$\mathrm{EVSI} = \mathbb{E}_{\boldsymbol{Y}}\left[\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left(\mathrm{NB}_d(\boldsymbol{\theta})\right)\right] - \max_d \mathbb{E}_{\boldsymbol{\theta}}\left(\mathrm{NB}_d(\boldsymbol{\theta})\right)$$

Where $\boldsymbol{Y}$ represents the data we plan to collect to inform our decision model.

As we mentioned in §1.3.2, there are several ways to compute EVSI. In this work, we consider the nonparametric regression-based method (Strong et al. (2015)). In this approach, multiple nonparametric regression equations (one for each possible treatment) are fitted to compute the preposterior mean:

$$\mathrm{NB}_d(\boldsymbol{\theta}) = g_d(T(\boldsymbol{Y})) + \epsilon$$

Where $\boldsymbol{\theta}$ is a sample of the joint distribution of the input parameters obtained by performing a probabilistic analysis (PA), and T is some low-dimensional summary statistic of the data as deepened in §1.3.2.

As mentioned above, previous examples of calculating EVSI have only generated $\boldsymbol{Y}$ from simple data collection schemes, which can be unrealistic in concrete scenarios. This work aims to introduce a methodology that enables us to compute the value of collecting more data when some of the existing data exhibit realistic missing mechanisms. To achieve this, we must first introduce the possible types of missing data that can occur and how we can simulate data with different types of missingness.

3.2.1 Missing data simulation

3.2.1.1 Simulate the full RCT data

First, we need to define how to simulate individual-level synthetic RCT data. In this context, we assume that the parameters governing the distribution of the relevant quantities (covariates, outcomes, and exposure variables) are derived from the literature or other forms of previous knowledge. Notice that in §2.2.2.1, a methodology to learn these parameters is provided in case we possess real data from an RCT (also a preliminary one). In this case, the RCT should satisfy some basic requirements to ensure that these parameters are informative for the simulation study. In particular, the RCT should present population, intervention, comparison and measured outcomes comparable to the target population of interest. In this case we would be able to infer the distribution of covariates and the relationships between covariates and outcomes from the RCT. Conversely, if we do not have an initial RCT that satisfies these conditions, relevant parameters can be informed using the existing literature. For example, published epidemiological studies, observational cohorts, meta-analyses or different literature evidence. Similarly, treatment effect estimates from clinical trials in related populations can be used to define plausible ranges or prior distributions for the relevant model components. In practice, evidence from multiple sources will often need to be combined, ensuring internal coherence between covariate distributions, outcome models, and expected treatment effects.

In this context, we introduce the following notation for a general RCT data with a population of n individuals. Assume that, each individual presents S covariates $\{X_1, X_2, .., X_S\}$, the exposure variable Z and the outcome variable Y.

To sample RCT data, we factorize the joint distribution of the covariates and the outcome variable in a product of sequential conditional distributions. In particular, for X_1, we first define the marginal distribution as $X_1 \sim F_1(\boldsymbol{\Theta}_0)$

We proceed by defining the conditional distribution of $X_2|X_1$ as $X_2|X_1 \sim F_2(\boldsymbol{\Theta}_1(X_1))$ and we go on recursively to define the other relevant distributions for the remaining covariates.

We assign Z randomly to each individual $Z_j \overset{iid}{\sim} Bern(q)$, for $j = 1, .., n$, where q is the expected proportion of treated individuals. Note that in the application, we will set the first nq individuals to be treated and the remaining $n(1-q)$ to be in the control group, as this is equivalent. Notice that the proportion of missing individuals has been reported to have a median of about 0.2, with some studies exceeding 0.6 (Fiero et al. (2016)). Given this variability and the risk of high missing percentage, especially in long-term studies or those relying on patient-reported outcomes, in concrete applications of the method introduced in this chapter, sensitivity analyses to explore different levels of missingness has to be performed.

Finally, for each individual, we define the distribution of the outcome given the covariates and the exposure variable as $Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$. In the EVSI computation, the outcome distribution depends on the PA simulations of the relevant parameters; this concept will be further explored in the next section.

Again, if we have access to real RCT data, a Bayesian nested regression model can be used to infer the relevant posterior distribution of the parameters and to create coherent RCT data as explained in §2.2.2.1.

We now have a simulation of RCT data $(\boldsymbol{Y}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$, the next step is to choose which type of missingness we want to reproduce. There are different types of missing

generating mechanisms, and for each of them, there is a way to induce it inside the data.

3.2.2 Simulating missingness

As we introduced in §1.2.3, there are several types of missing mechanisms we may encounter in real scenarios (Rubin (1976), Little and Rubin (2019)).

3.2.2.1 Missing completely at random

In a *missing completely at random* (MCAR), the probability that an individual misses does not depend on the covariates $\boldsymbol{X}$ nor on the outcome Y.

In this case, to simulate MCAR data, one has to sample the missingness indicator π_j for each of the n individuals, $\pi_j \sim Bern(p_0)$ for $j = 1, ..., n$, with p_0 the expected percentage of missingness that we plan to see in the data. Finally, one has to set $Y_j = \text{NA}$ for those individuals having $\pi_j = 1$. This is the simplest missing mechanism, but also the most unrealistic one since, in real-world scenarios, there is nearly always an interaction between the individuals' characteristics and the probability one has to be missing in the related study.

3.2.2.2 Missing at random

Differently, in a *Missing at random* (MAR) model (Rubin (1987)), the missing generating mechanism depends on the covariates $\boldsymbol{X}$ but not on the outcome Y.

In this scenario, we can simulate missing data by first defining the individual probability of missingness p_j for $j = 1, \ldots, n$ as follows:

$$p_j = G_{\text{MAR}}(\beta_{00} + G^{-1}_{\text{MAR}}(p_0) + g_{\text{MAR}}(\boldsymbol{\beta}, \boldsymbol{X}_j))$$

where p_0 is the expected percentage of missingness. In a realistic scenario, $\boldsymbol{\beta}$ would be coefficients informed by the literature, g represents how the probability of missingness

depends on the covariates $\boldsymbol{X}_j$, and G^{-1} is the link function of this generalized binary model. Note that for g_{MAR} linear and G_{MAR} the CDF of a Normal distribution, we recover the probit model (Bliss (1934)). In contrast, for G_{MAR} the CDF of a logistic distribution, we recover the standard logit model (Berkson (1944)). Note that in practice, informing $\boldsymbol{\beta}$ from the literature requires studies reporting predictors of missingness (e.g., loss to follow-up or non-response associated with patient characteristics) or relevant experts' opinions. Such information is sometimes available but often incomplete, making literature-based estimates feasible yet approximate. For this reason, using plausible ranges for $\boldsymbol{\beta}$ and conducting sensitivity analyses is generally recommended.

The parameter β_{00} is estimated with numerical methods so that the following equation is satisfied:

$$\frac{1}{n}\sum_{j=1}^{n} G_{\text{MAR}}(\beta_{00} + G^{-1}_{\text{MAR}}(p_0) + g_{\text{MAR}}(\boldsymbol{\beta}, \boldsymbol{X}_j)) - p_0 = 0$$

Note that this ensures that the expected percentage of missing outcomes in the generated data is p_0. This is important since we want the missing proportion to match the initial expectation.

As above, we finally sample the missingness indicator π_i for each of the n individuals, $\pi_j \sim Bern(p_j)$ for $j = 1, ..., n$, and set $Y_j = \text{NA}$ for those individuals having $\pi_j = 1$.

3.2.2.3 Missing not at random

Finally, we may face a *Missing not at random* (MNAR) (Molenberghs and Kenward (2007), Schafer and Graham (2002)) model, when the missing generating mechanism depends on both the covariates $\boldsymbol{X}$ and the outcome Y;

To reproduce an MNAR model, we define p_j as:

$$p_j = G_{\text{MNAR}}(\beta_{00} + G^{-1}_{\text{MNAR}}(p_0) + g_{\text{MNAR}}(\boldsymbol{\beta}, \boldsymbol{X}_j, Y_j))$$

where, this time, g_{MNAR} represents how the probability of missingness depends on the

covariates $\boldsymbol{X}_j$ and the outcome Y_j, and G_{MNAR} as above. In this case, since the MNAR mechanism is not identifiable, the parameters governing the dependence on Ycannot be estimated empirically and are instead informed by prior knowledge, experts' opinions, or qualitative evidence on likely dropout mechanisms. As recommended in the relevant literature (Molenberghs and Kenward (2007)), these assumptions should be explored through sensitivity analyses over a range of plausible values."

β_{00} is estimated with numerical methods so that the following equation is satisfied:

$$\frac{1}{n}\sum_{i=1}^{n} G_{\text{MNAR}}(\beta_{00} + G^{-1}_{\text{MNAR}}(p_0) + g_{\text{MNAR}}(\boldsymbol{\beta}, \boldsymbol{X}_j, Y_j)) - p_0 = 0$$

again, note that this ensures that the expected percentage of missing outcomes in the generated data is π_0;

Once we generate the probability of a missing outcome p_j for each individual j we can generate the missingness indicator as follows:

$$\pi_j \sim Bern(p_j) \quad j = 1, .., n$$

and for the individuals who exhibit $\pi_j = 1$, we set $Y_j = \text{NA}$.

As said, MCAR models are unrealistic. For this reason, this work will focus solely on MAR and MNAR models. With this methodology, we can generate RCTs with missing observations, a common situation in real-world scenarios.

In realistic settings, when analyzing data with missing observations, we typically have several choices (Gabrio et al. (2017)), as highlighted in §1.2.3. First, we can work with complete data, discarding all individuals with missing outcomes. This method typically yields inaccurate and/or biased estimates of the quantities of interest, due to both analyzing a smaller amount of data (which increases variance in the estimate) and overlooking the fact that missingness can be associated with specific relevant characteristics of individuals. Second, by assuming a MAR missing mechanism and estimating the dependence

between the outcome and covariates using different techniques (OLS, GLM, GAM, etc.), one can perform multiple imputation methods §1.2.3 (Rubin (1987)). Lastly, if we assume that the underlying missingness mechanism is MNAR, we must model the missingness explicitly, jointly modeling the missing indicator variable and the outcome.

We can now proceed to introduce the general methodology for computing EVSI in this context.

3.3 VoI analysis on missing data

When dealing with missing data, the process in §1.3.1 and §1.3.2 must be modified. If we plan to collect data with missingness $(\boldsymbol{Y}_m, \boldsymbol{X}, \boldsymbol{Z})$, then we must modify the standard process above by sampling those kinds of data instead of idealized RCTs. This step will be done by first sampling full RCT data as introduced in §3.2.1.1, and then by simulating missingness as presented in §3.2.2.

Once data affected by missingness are sampled, we have successfully reproduced the realistic situation in which a researcher collects additional *realistic* RCTs, and is aiming to determine whether there is value of information of the sampled data.

In a realistic scenario, before the analysis and the economic evaluation, one has to impute the missing data using different techniques, as mentioned. Similarly, when using non-parametric regression methods to compute EVSI, we cannot directly use $T(Y_m)$ in our estimates. Therefore, we must first impute missing data to find an unbiased estimate of $T(Y_{\text{RCT}})$.

We denote as $(\boldsymbol{Y}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ the simulated individual-level RCT data, simulated at the beginning of the process. It represents a dataset composed of n individuals with outcomes $\boldsymbol{Y}_{\text{RCT}}$, covariates $\boldsymbol{X}$ and exposures variables $\boldsymbol{Z}$. We also denote as $(\boldsymbol{Y}_m, \boldsymbol{X}, \boldsymbol{Z})$, RCT data with n individuals, obtained by multiple imputations, and as $\tilde{T}(\boldsymbol{Y}_m)$ the unbiased estimate of $T(\boldsymbol{Y}_{\text{RCT}})$.

Starting from the same definition of the underlying economic model and the net benefit, we now present our method for computing EVSI.

3.3.1 Simulating the covariates of interest

The first step of the novel methodology is to generate covariates $\boldsymbol{X}$ using the method described in §3.2.1.1. These covariates will be fixed throughout the first steps of EVSI calculation, meaning that initially, the data we generate will have different outcomes, but the same covariate profile. Covariates would be marginalized out only in the final step of the methodology to obtain unconditional EVSI. Here, we also generate the exposure variable Z from $Z_j \overset{iid}{\sim} Bern(0.5)$, for $j = 1, .., n$, so that allocation to different groups is entirely random. Note again that in applications, this is equivalent to choosing a priori which indexes represent treated and control individuals (e.g., the first m to the treatment group and the remaining $n - m$ to the control).

3.3.2 Simulating the patient-level outcomes

The next step is to obtain K samples of the relevant parameters from PA simulations $\boldsymbol{\theta}_i|Z \sim p(\boldsymbol{\theta}_0(Z_i))$, $i = 1, \ldots, K$ with p prior distribution of $\boldsymbol{\theta}$ governed by parameters $\boldsymbol{\theta_0}$. These parameters are the parameters we assume can be informed by the new data we want to collect; for example, they can be the mean, the rate, or another relevant measure of the outcome. Note that they have different distributions for different values of the exposure variable; this is because we expect the relevant parameter affecting the outcome of an individual to depend on the treatment status of the individual themselves as well. Now, since we already have sampled $\boldsymbol{\theta}$ from the PA simulation, before sampling the outcome Y, we want to impose dependence between Y, Z, and $\boldsymbol{X}$ by transforming the PA parameters to incorporate the above dependence:

$$\tilde{\boldsymbol{\theta}}_i = f(\boldsymbol{\theta}_i, \boldsymbol{X})$$

for instance we may have that $\tilde{\boldsymbol{\theta}}_i$ is a linear combination of $\boldsymbol{\theta}_i$ and $\boldsymbol{X}$:

$$\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \sum_{j=1}^{K} \beta_j (X_j - \overline{X})$$

or we can immediately incorporate dependence by sampling the PA simulations from a distribution whose parameters depend on the values of the covariates:

$$\tilde{\boldsymbol{\theta}}_i \sim p(\boldsymbol{\theta}_0, \boldsymbol{X}, Z) \quad i = 1, .., K$$

instead of

$$\boldsymbol{\theta}_i | Z \sim p(\boldsymbol{\theta}_0(Z_i)) \quad i = 1, .., K$$

This is important to make the sampled outcome dependend from the covariate profile.

The next step is to sample the outcome $\boldsymbol{Y}_{\text{RCT}}$ from $\boldsymbol{Y}_{RCT}^{(i)} \sim F(\tilde{\boldsymbol{\theta}}_i)$ and obtain K starting RCTs $(\boldsymbol{Y}_{\text{RCT}}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ were fixed at the beginning of the process.

3.3.3 Generating data with missingness and compute EVSI

Once we have sampled the synthetic RCT, we must decide the underlying missing mechanism and generate missing data as explained in §3.2.2 to obtain $(\boldsymbol{Y}_m^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$. Once the RCT with missing data is sampled, we can apply multiple imputation to impute data and compute $\tilde{T}(\boldsymbol{Y}_m)$, an unbiased estimator of $T(\boldsymbol{Y}_{\text{RCT}})$.

The last two steps are: first, estimate the preposterior mean with non-parametric regression methods as explained in §1.3.2 and obtain an estimate of EVSI$|\boldsymbol{X}$, as we fixed $\boldsymbol{X}$ at the beginning of the process. Then, finally, integrate the dependence on $\boldsymbol{X}$ with

MC computation and obtain EVSI:

$$\text{EVSI} = \mathbb{E}_{\boldsymbol{X}}(\text{EVSI}|\boldsymbol{X}) \approx \frac{1}{N}\sum_{i=1}^{N}(\text{EVSI}|\boldsymbol{X}_i)$$

This translates to performing the entire process N times, each time sampling different covariate profiles.

We now outline the key passages of the methodology described above.

1. Generate covariates $\boldsymbol{X}$ and the exposure variables $\boldsymbol{Z}$ for n individuals $j = 1, .., n$.

2. Sample K parameters $\boldsymbol{\theta}_i$, $i = 1, \ldots, K$ inside the PA simulations.

3. Develop the individual-level parameters $\tilde{\boldsymbol{\theta}}_i$ required to simulate the outcomes, inducing dependence of the relevant parameters with covariates.

4. Sample new data (the outcome or another relevant measure) governed by the relevant parameters, as in the standard approach, coming from an RCT

$$\boldsymbol{Y}_{\text{RCT}}^{(i)} \sim F(\tilde{\boldsymbol{\theta}}_i)$$

   and obtain K starting RCTs $(\boldsymbol{Y}_{\text{RCT}}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ and $\boldsymbol{Z}$ have been fixed at the beginning of the process.

5. Decide the underlying missing mechanism and generate missing data as explained in 3.2.2.

6. Once we have $(\boldsymbol{Y}_m^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ we can apply multiple imputation to impute data and compute $\tilde{T}(\boldsymbol{Y}_m)$, an unbiased estimator of $T(\boldsymbol{Y}_{\text{RCT}})$.

7. Estimate preposterior mean with nonparametric regression methods and obtain an estimate of EVSI$|\boldsymbol{X}$.

8. Finally integrate out with MC computation the dependence on $\boldsymbol{X}$ and obtain the final estimate of the EVSI.

3.4 Application

In this section, we will apply the methodology described above in two different contexts. In the first one, we assume a Normal-Normal model and apply the methodology in a simulation-based scenario in which we can control the effect of each parameter of the health economic model. In the second one, we consider a realistic scenario with an underlying Markov health-economic model

3.4.1 Normal-Normal model

Starting from a Normal-Normal model, we recall that our objective is to generate one data set with missingness for each of the K PA simulation and to use those data to compute the EVSI. We define $Z \in \{0, 1\}$ as the exposure variable and assume to know a priori how many people belong to each group. We set n_0 as the number of people in the control group and n_1 as the ones in the treatment group. In this application, we also assume that $n_0 = n_1$.

We set the prior for the mean outcome in the treatment and control groups, respectively, as a normal distribution with means θ_1 and θ_0, and variances σ_1^2 and σ_0^2. Thus, formally, the priors for the mean outcome, conditional on Z, are:

$$\theta_{Z=0}^{(i)} = \theta^{(i)} | Z = 0 \sim N(\theta_0, \sigma_0^2) \quad i = 1, .., K \tag{3.1}$$

$$\theta_{Z=1}^{(i)} = \theta^{(i)} | Z = 1 \sim N(\theta_1, \sigma_1^2) \quad i = 1, .., K. \tag{3.2}$$

We assume the the data collection consists of $n_1 = n_2$, with $n_1 + n_2 = n$, independent

samples from:

$$Y_j^{(i)}|Z=0 \sim N\left(\theta_{Z=0}^{(i)}, \sigma_{\boldsymbol{Y}}^2\right) \quad j=1,...,n_1 \quad i=1,..,K \tag{3.3}$$

$$Y_j^{(i)}|Z=1 \sim N\left(\theta_{Z=1}^{(i)}, \sigma_{\boldsymbol{Y}}^2\right) \quad j=1,...,n_2 \quad i=1,..,K \tag{3.4}$$

with $\sigma_{\boldsymbol{Y}} = 20$.

The net benefit functions for this model are given by:

$$\begin{aligned} NB_0(\theta) &= k\theta_0 \\ NB_1(\theta) &= k\theta_1 - c \end{aligned} \tag{3.5}$$

Given these ingredients, we can easily compute the EVSI for a given sample size n by constructing PA simulations from (3.1) and (3.2) and using (3.3) and (3.4) as the data generating process.

Starting from this, we apply the methodology described in the previous section and compute EVSI with missing data, then compare it with the original EVSI in the normal-normal model. We refer to the last section, following each of the steps defined above, to illustrate how the general methodology applies in the context of the Normal-Normal model introduced earlier. “Note that we made the number of each passage corresponding with the ones in the final scheme of §3.3 for a better understanding. First, we assume that we will collect data on a set of $n = 100$ individuals, each with a different *age* and a different number of comorbidities *com*, and measure the effectiveness Y of a specific therapy in these individuals.

1. We simulate 2 independent covariates for each individual j: *age, com* and we assume that the first half of the individuals are treated ($Z = 1$), while the others belong to the control group ($Z = 0$). We set $n_0 = 50$ and $n_1 = 50$, respectively, the number

of people in the control and treatment group, and $n = n_1 + n_2$.

$$age_j \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{age}, \sigma^2_{age}) \quad j = 1, ..., n$$

$$com_j \overset{\text{iid}}{\sim} Bin(n_{com}, p_{com}) \quad j = 1, ..., n$$

with $\mu_{age} = 57$, $\sigma_{age} = 8$, $n_{com} = 10$ and $p_{com} = 0.5$.

2. First, we sample for each of the $K = 5000$ PA simulations, θ:

$$\theta^{(i)}_{Z=0} = \theta^{(i)}|Z = 0 \sim N(\theta_0, \sigma^2_0) \quad i = 1, .., K$$

$$\theta^{(i)}_{Z=1} = \theta^{(i)}|Z = 1 \sim N(\theta_1, \sigma^2_1) \quad i = 1, .., K$$

with $\theta_0 = 20$, $\theta_1 = 40$, $\sigma_1 = 8$ and $\sigma_0 = 6$.

3. For each simulated individual j, we compute $\tilde{\theta}_i$ (the mean of the outcome distribution, given the patient's i characteristics) as follows:

$$\tilde{\theta}^{(i)}_j = \theta^{(i)}_{Z_j} + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) \quad j = 1, ..., n$$

where $\theta^{(i)}_{Z_j}$ is the i-th sample inside PA simulation for group Z_j and $\alpha_1 = 3, \alpha_2 = 5$ are coefficients that would have been derived from the literature in a realistic scenario.

4. We sample the outcome for individual j as:

$$Y^{(i)}_j \sim N(\tilde{\theta}^{(i)}_j, \sigma^2_Y)$$

and obtain $K = 5000$ starting RCTs $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ were fixed at the beginning of the process.

5. We set $\pi_0 = 0.5$ as the expected percentage of missingness in the data and define the

missing mechanism (MAR or MNAR). For each simulated individual j, we compute p_j (the probability of a missing outcome) as follows:

- In a MAR model we define p_j as:

$$p_j = \Phi(\beta_{00} + \Phi^{-1}(p_0) + \beta_1(age_j - \overline{age}) + \beta_2(com_j - \overline{com}))$$

where $\beta_1 = 0.06, \beta_2 = 0.1$ would be coefficients informed by the literature in a realistic scenario. β_{00} is estimated with numerical methods so that the following equation is satisfied:

$$\frac{1}{n}\sum_{j=1}^{n} \Phi(\beta_{00} + \Phi^{-1}(p_0) + \beta_1(age_j - \overline{age}) + \beta_2(com_j - \overline{com})) - p_0 = 0$$

Note that this ensures that the expected percentage of missing outcomes in the generated data is p_0;

- In a MNAR model we define p_j as:

$$p_j = \Phi(\beta_{00} + \Phi^{-1}(p_0) + \beta_1(age_j - \overline{age}) + \beta_2(com_j - \overline{com}) + \beta_3(Y_j - \overline{Y}))$$

where $\beta_1 = 0.06, \beta_2 = 0.1, \beta_3 = 0.2$ are the coefficients, typically informed by the literature in a realistic scenario.

β_{00} is estimated with numerical methods so that the following equation is satisfied:

$$\frac{1}{n}\sum_{j=1}^{n} \Phi(\beta_{00} + \Phi^{-1}(p_0) + \beta_1(age_j - \overline{age}) + \beta_2(com_j - \overline{com}) + \beta_3(Y_j - \overline{Y})) - p_0 = 0$$

again, note that this ensures that the expected percentage of missing outcomes in the generated data is p_0.

Once we have the probability of a missing outcome p_j for each individual j we sample the missingness indicator as follows:

$$\pi_j^{(i)} \overset{\text{iid}}{\sim} Bern(p_j) \quad j = 1, .., n \quad i = 1, ..., K$$

and for the individuals that exhibit $\pi_j = 1$ we set $Y_j = \text{NA}$

6. Once we obtain K simulated data with missingness $(\boldsymbol{Y}_m^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, ..., K$. For each of the latter, we perform multiple imputation, using GLM and imputing the dataset 20 times, to obtain $\tilde{T}(\boldsymbol{Y_m})$, an unbiased estimator of $T(\boldsymbol{Y}_{\text{RCT}})$. T in this case is the sample mean, $T(\boldsymbol{Y}) = \overline{\boldsymbol{Y}}$ and it is a sufficient statistic of the parameters $\theta^{(i)}$, a desirable property for T, as specified in §1.3.2.

7. We estimate the preposterior mean from the following regression with Generalized additive model (GAM):
$$NB(\boldsymbol{\theta}_i) = g(\tilde{T}(\boldsymbol{Y}_m)) + \epsilon \tag{3.6}$$
and obtain an estimate of $EVSI|\boldsymbol{X}$;

8. Finally, we integrate out the dependence on $\boldsymbol{X}$ and obtain the final estimate of the EVSI as:
$$EVSI = \frac{1}{N} \sum_{i=1}^{N} EVSI|\boldsymbol{X}_i$$
This is achieved by computing $EVSI|\boldsymbol{X}$ N times, where a different covariate profile is sampled each time, and then averaging over the obtained EVSI curves. In this and the next application, we set $N = 10$, as numerical results indicate that this is sufficient, given the low variance between EVSI curves obtained from different covariate profiles.

We will show the results of this procedure and why it is important to take into account

the realistic nature of the data we plan to collect in the next section. Now, we introduce the second relevant application that represents a more realistic scenario.

3.4.2 Chemotherapy Markov treatment model

We briefly recall the decision model introduced in §1.1.2. The model combines two common structures: a decision tree model and a Markov state transition model. In the decision tree model, we aim to estimate the proportion s_d of individuals who experience side effects after receiving standard care, $d = 0$, or novel treatment, $d = 1$. While the Markov model represents the possible trajectories of the health state and their related costs that individuals with side effects might incur.

In a standard scenario, we would simulate future outcomes s_0 and s_1 and then apply the usual procedure to compute EVSI. In this novel context, we follow the general methodology introduced in §3.3 to compute EVSI in the presence of missing data. In this case, the only differences with respect to the previous example are: 1) the different distributions of PA, 2) the outcome distribution, and 3) the statistic T is no longer the sample mean of the outcomes, but rather the odds ratio. Moreover, in this application, we consider a sample size of $n = 1000$. Therefore, the first four steps of the general methodology become the following:

1. We simulate two independent covariates for each individual j: *age* and *com*, and we assume that the first half of the individuals are treated and the other half are not. We set $n_0 = 500$ and $n_1 = 500$, respectively, the number of people in the control and treatment group, and $n = n_1 + n_2$.

$$age_j \overset{\text{iid}}{\sim} \mathcal{N}(57, 64) \quad j = 1, ..., n$$

$$com_j \overset{\text{iid}}{\sim} Bin(10, 0.5) \quad j = 1, ..., n$$

2. Then, we sample $K = 5000$ PA simulations with relevant parameters sampled from

$$log(\rho^{(i)}) \sim N(log(0.54), 0.3^2) \quad i = 1, .., K$$

$$s_0^{(i)} \sim Beta(1, 1) \quad i = 1, .., K$$

$$s_1^{(i)} = 1 - \frac{1 - s_0^{(i)}}{s_0^{(i)}(\rho^{(i)} - 1) + 1} \quad i = 1, .., K$$

and all the other relevant parameters listed in §1.1.2. ρ is the odds ratio of a side effect occurrence, s_0 is the prior probability of a side effect occurrence in the control group, while s_1 is the analogous for the treated group. For a complete list of the parameters and their distributions, we refer to §1.1.2 and (Heath et al. (2024)).

3. For each simulated individual j, we compute $\tilde{s}_j^{(i)}$ (the mean of the outcome distribution, that is the probability for individual j of having a side effect, given the patient's j characteristics) as follows:

$$\tilde{s}_{0,j}^{(i)} = \Phi(\Phi^{-1}(s_0^{(i)}) + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) + \alpha_{0,0}) \quad j \text{ s.t. } Z_j = 0$$

where $s_0^{(i)}$ is the i-th sample of PA simulation (for control group), $\alpha_1 = 0.05$, $\alpha_2 = 0.3$, Φ is the CDF of a standard Gaussian and $\alpha_{0,0}$ is such that:

$$\sum_{j=1}^{n} \Phi(\Phi^{-1}(s_0^{(i)}) + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) + \alpha_{0,0}) - s_0^{(i)} = 0$$

so that the individual parameters preserve the overall mean $s_0^{(i)}$. In an analogous way we define:

$$\tilde{s}_{1,j}^{(i)} = \Phi(\Phi^{-1}(s_1^{(i)}) + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) + \alpha_{0,1}) \quad j \text{ s.t. } Z_j = 1$$

4. We sample the outcome for individual j as:

$$Y_j^{(i)} \overset{\text{iid}}{\sim} Bern(\tilde{s}_j^{(i)})$$

and obtain $K = 5000$ starting RCTs $(\boldsymbol{Y}_{RCT}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ were fixed at the beginning of the process.

Then we proceed in an almost identical way to the Normal Normal model with the only significant difference that, this time, the statistics T will be the odds ratio of side effects, and the parameters to inform in the EVSI process are s_0 and s_1.

3.4.3 Comparison and second moment matching

In the next section, we will present our results by comparing the following objects:

1. Original EVSI, equal to EVSI computed on RCT data with no missingness, that is EVSI calculated relying on the $K = 5000$ starting RCTs $(\boldsymbol{Y}_{\text{RCT}}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$.

2. EVSI computed on complete cases, that is, on fully observed data.

3. EVSI computed on data obtained through the multiple imputation method under the MAR assumption. Note that by assuming MAR, multiple imputation gives unbiased estimates of the relevant measures.

4. EVSI computed on data obtained through the multiple imputation method under the MNAR assumption. With respect to the previous scenario, multiple imputation returns biased estimates under MNAR assumptions as the missing mechanism also depends on partially observed variables.

We will show in the next section that the EVSI computed on data with missing observations will be lower than the original EVSI. Given that the higher the sample

size of simulated data, the higher the EVSI (Heath et al. (2019)), we propose a method to estimate the sample size required to reach the initial level of information under the assumption of an MAR missing mechanism. This method works by matching the second moment of $\tilde{T}(\boldsymbol{Y}_m)$ to the one of the original $T(\boldsymbol{Y}_{\text{RCT}})$ for an increasing sample size of $(\boldsymbol{Y}_m, \boldsymbol{X}, \boldsymbol{Z})$, as by doing this, we will match not only the first (already matched as $\tilde{T}$ unbiased) but also second moments of the regressors in the nonparametric regression (3.6), obtaining therefore close preposterior distributions. We will further develop this method and present the results in the following section.

3.5 Results

3.5.1 Normal-Normal missing model

As we specified in the previous section, to show the results in the Normal-Normal context, we start by comparing the following quantities:

1. Original EVSI (equal to EVSI computed on RCT data with no missingness);

2. EVSI computed on data obtained through multiple imputations method (under MAR assumption).

The plot in Figure 3.1 represents the above quantities under the parameter values defined in §3.4.1.

We can immediately notice that the EVSI value decreases when the data generated are affected by missingness. This is mainly due to the additional variability introduced by the missingness mechanism and the multiple imputation within the data-generating mechanism.

The first source of variability is related to the fact that before collecting additional data, we do not know which data will be missing, and we must account for this additional

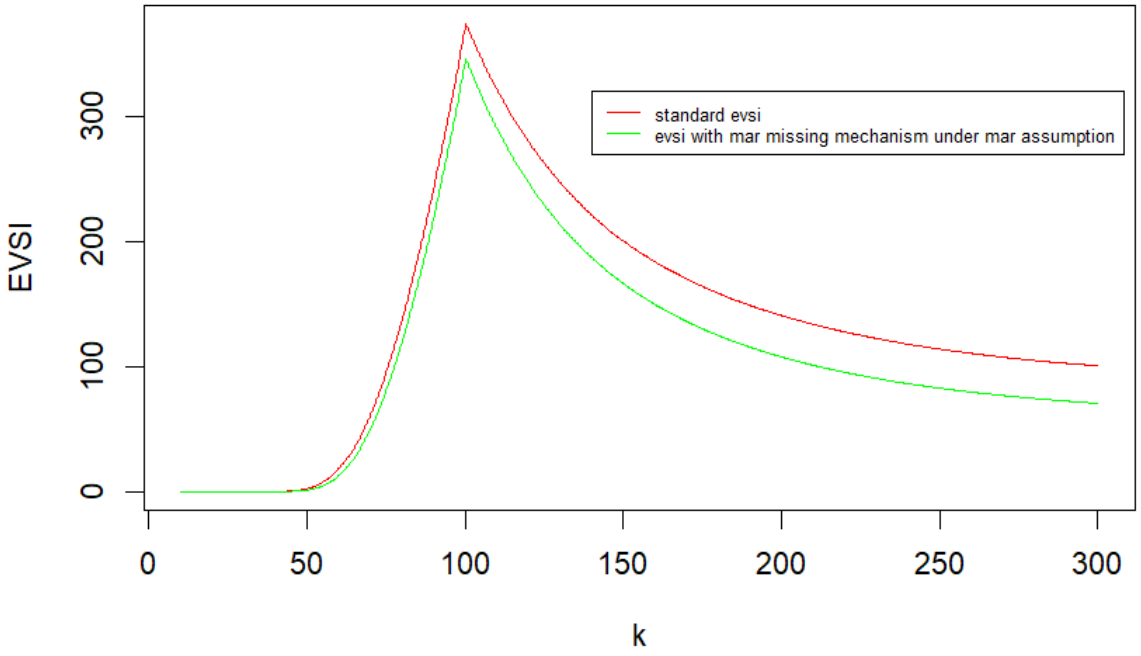


Figure 3.1: EVSI calculated relying on the $K = 5000$ starting RCTs $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$ in red and EVSI computed on data obtained through multiple imputations $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MAR assumption in green.

uncertainty in our model. The second is related to the additional variance in the estimations that the multiple imputation method adds to the general model. These two layers of additional variance must be taken into account when we plan to collect more realistic data.

As we mentioned in the previous section, it can also be useful to compare these results with EVSI on fully observed data, i.e., excluding individuals with missing outcomes from the analysis without using any techniques to address the missingness bias. The result is shown in Figure 3.2.

As we observe, the fact that multiple imputations produce a complete dataset compensates only in part for the loss in precision that we incur when considering only the complete case analysis. In other scenarios, the loss in precision due to the two sources of

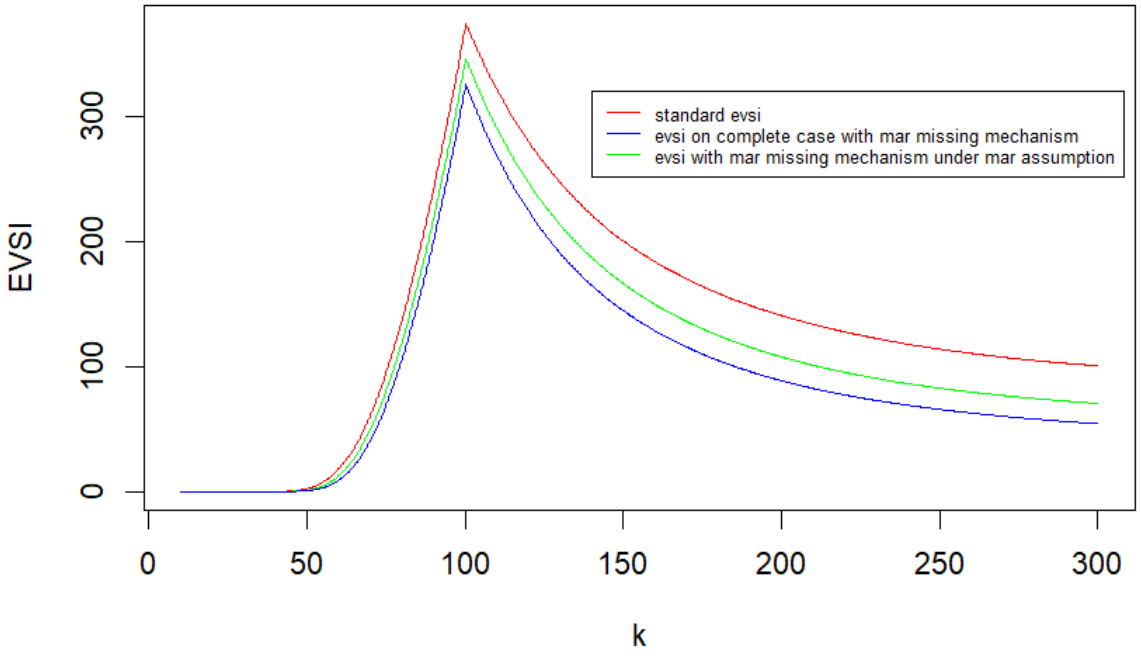


Figure 3.2: EVSI calculated relying on RCT data $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$ in red, EVSI computed on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MAR assumption in green, EVSI on fully observed data (complete case) in blue

uncertainty above can be even greater than that of the complete case analysis due to the greater variability of the different estimates.

Let us now compare, in Figure 3.3, the scenario above with a context in which we face an MNAR missing mechanism.

As we can observe, model misspecification and the related biased estimates translate into a decreasing level of EVSI. This behaviour must be further investigated to better understand the effects of an unbiased estimate of $T(\boldsymbol{Y}_{\text{RCT}})$ in the EVSI computation.

As we mentioned in the previous section, in the case of a MAR missing mechanism, there is a natural way to adjust for the additional sources of uncertainty introduced by missingness within the model in RCT analysis, namely by increasing the sample size of the data we plan to collect. The straightforward way to compute the necessary sample size to

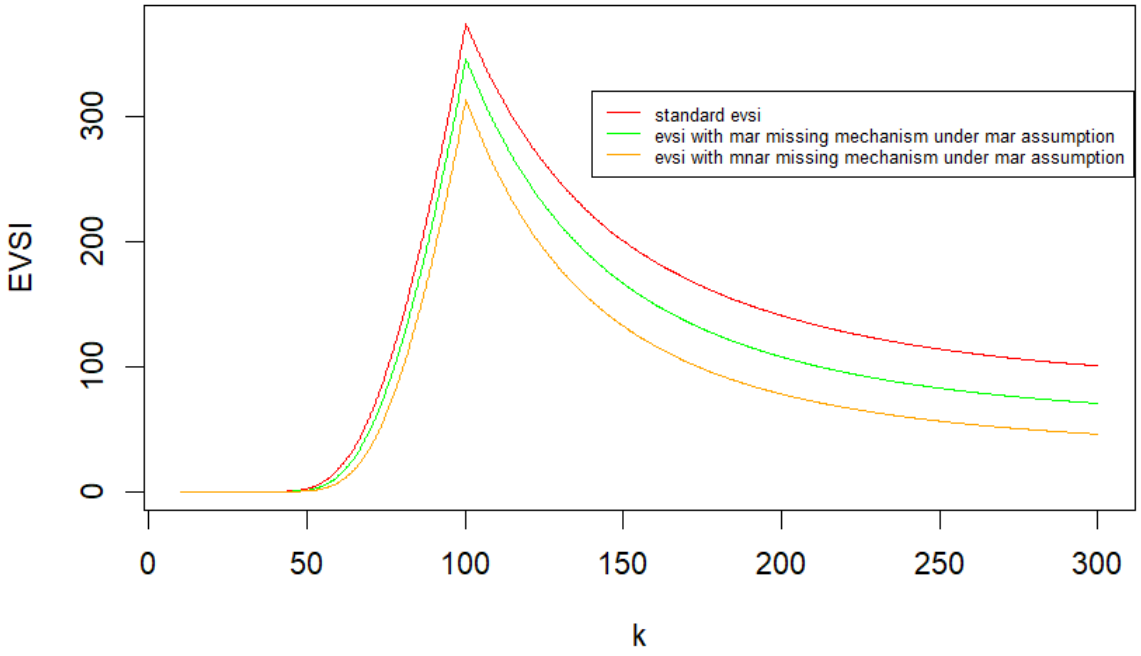


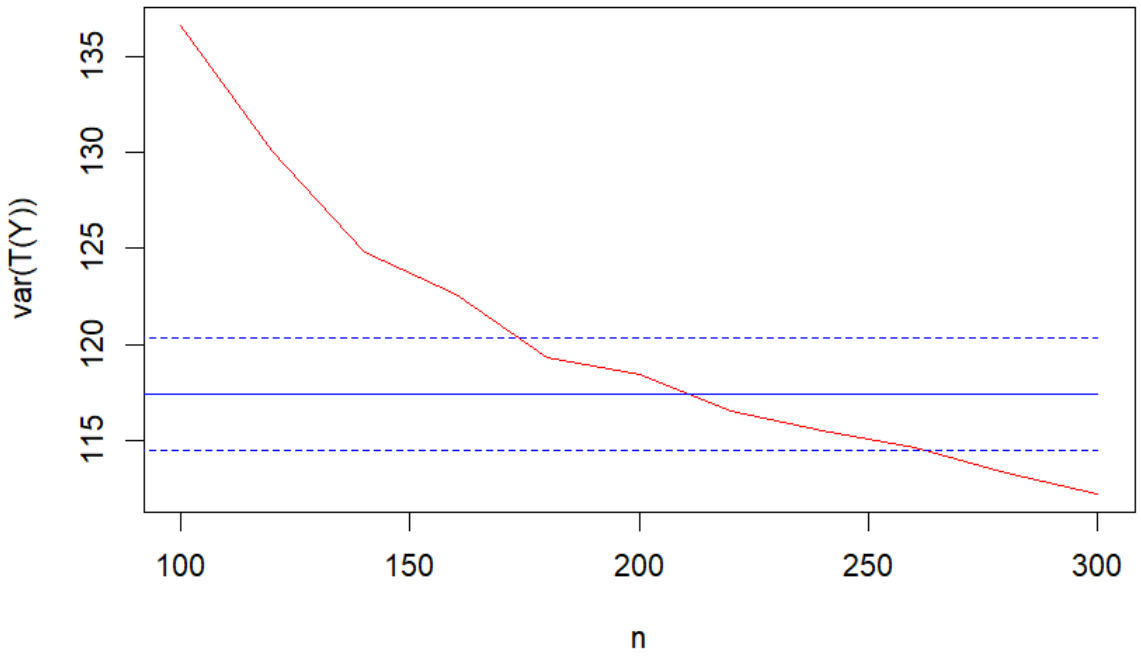
Figure 3.3: EVSI calculated relying on RCT data $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$ in red, EVSI computed on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MAR assumption in green, EVSI on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MNAR assumption in orange

achieve the original level of EVSI is to sequentially increase the sample size in the EVSI computation until the original level is reached; nevertheless, this can be computationally expensive.

One way to properly take into account the increase in variability due to missingness is to compute the variance of $\tilde{T}(\boldsymbol{Y}_m)$ (under an MAR missing mechanism) and to calculate for which value of the sample size n it equals the variance of $T(\boldsymbol{Y}_{RCT})$ in the original EVSI, that is, when there is no missingness inside the data. Since we know that multiple imputation methods produce unbiased estimates of $T(Y)$, by matching the variance as well, we can reach the previous level of information.

In the plot in Figure 3.4, the horizontal blue line represents the value of $\text{var}(T(\boldsymbol{Y}_{RCT}))$, while the dashed lines represent the bounds of a symmetric interval $(var(T(\boldsymbol{Y}_{RCT})) -$

$0.025var(T(\boldsymbol{Y}_{RCT})), var(T(\boldsymbol{Y}_{RCT})) + 0.025var(T(\boldsymbol{Y}_{RCT}))).$



.

Figure 3.4: Comparison between $\text{var}(\tilde{T}(\boldsymbol{Y}_m))$ for increasing sample size n and $\text{var}(T(\boldsymbol{Y}_{\text{RCT}}))$

Finally, considering the previous graph, we estimate an adjusted sample size $n_{adj} \approx 210$ and calculate the EVSI of the data affected by missingness for this particular sample size. The result is shown in Figure 3.5.

As we can see, the original level of EVSI has finally been reached. Note that the original sample size was $n = 100$ with a percentage of missingness $p_0 = 0.5$, and the adjusted sample size to match the variance of $T(\boldsymbol{Y}_{RCT})$ has been estimated in $n_{adj} \approx 210$. Typically, in studies affected by missingness, the adjustment researchers make follows this relation $n_{adj} \approx n/(1 - p_0)$, meaning that using a standard approach to sample size calculation in this specific scenario, we would have collected data with a sample size of 200,

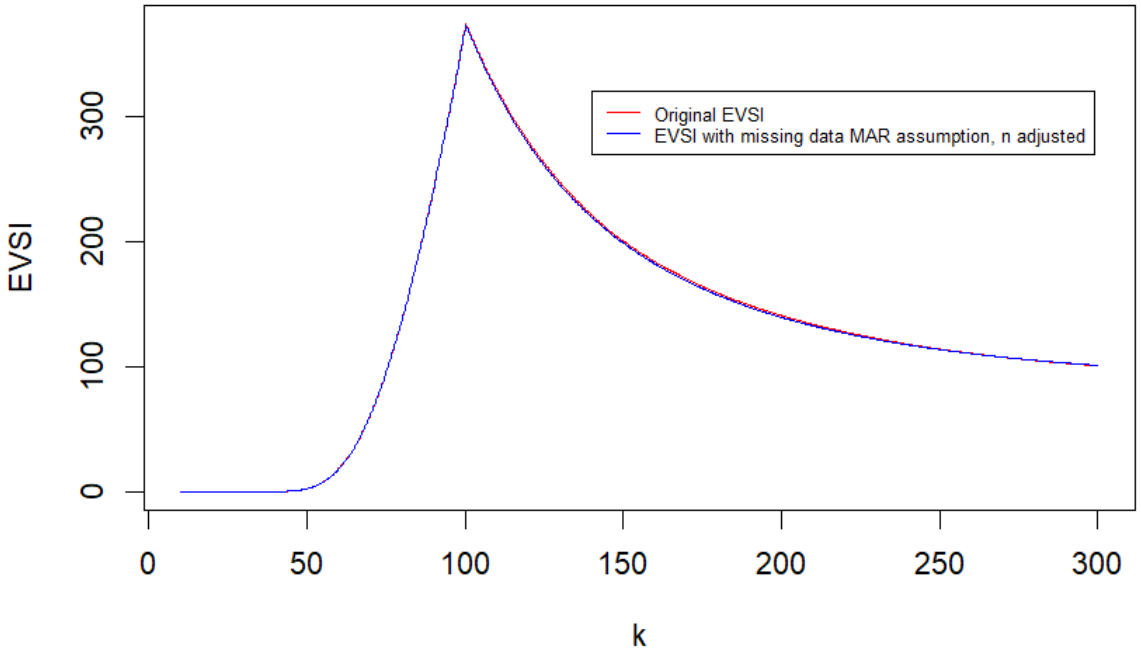


Figure 3.5: EVSI calculated relying on RCT data $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$ in red, EVSI computed on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MAR assumption with adjusted sample size in blue

with the consequence of not reaching the original level of EVSI. This deviation from the standard approach to account for missing data will be even more evident in the following application.

3.5.2 Chemotherapy treatment model

We now applied the above methodology to the more realistic context of the Markov decision model introduced in §1.3.3.1.

The plot in Figure 3.6 represents the original EVSI (without missingness), the EVSI on complete data, the EVSI with MAR missing mechanism, and the EVSI with MNAR missing mechanism.

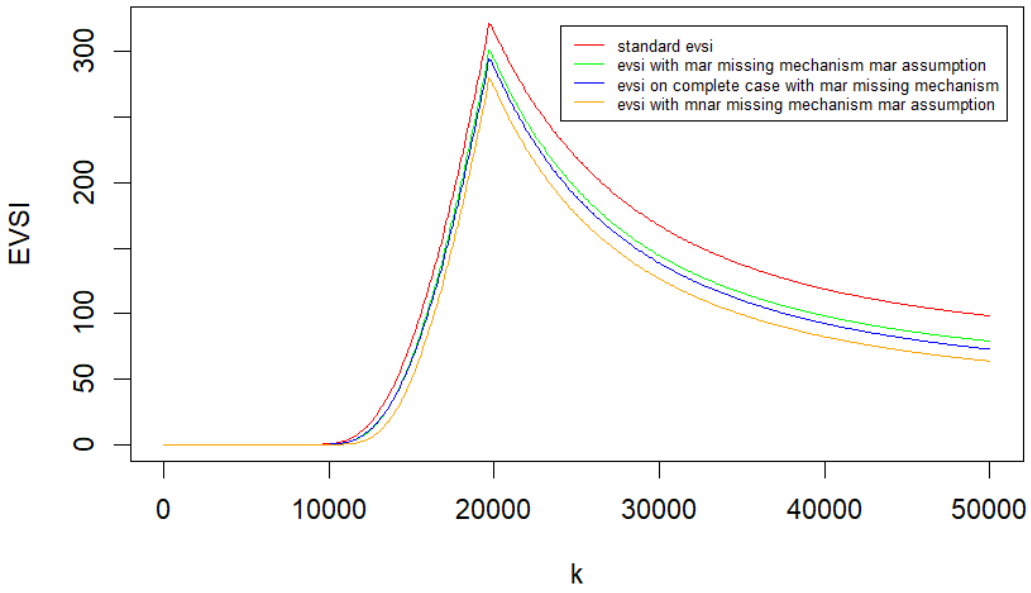
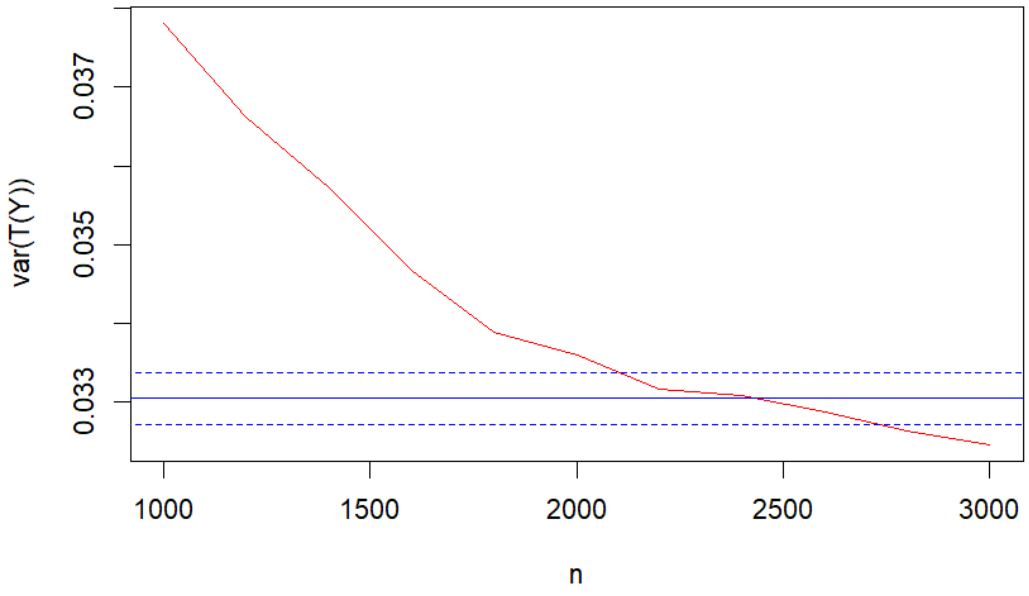


Figure 3.6: EVSI calculated relying on RCT data $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$ in red, EVSI computed on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MAR assumption in green, EVSI on fully observed data (complete case) in blue, EVSI on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MNAR assumption in orange

We then proceed as we did in the Normal conjugate model application taking the bias into account correctly by computing the variance of $\tilde{T}(\boldsymbol{Y}_m)$ (under an MAR missing mechanism) and calculating for which value of the sample size n it equals the variance of the original $T(\boldsymbol{Y}_{RCT})$, without missingness. This is represented in Figure 3.7.

Finally, considering the previous graph, we set $n \approx 2450$ and compute the EVSI of the data affected by missingness for this precise sample size, as shown in Figure 3.8.



.

Figure 3.7: Comparison between $\text{var}(\tilde{T}(\boldsymbol{Y}_m))$ for increasing sample size n and $\text{var}(T(\boldsymbol{Y}_{\text{RCT}}))$

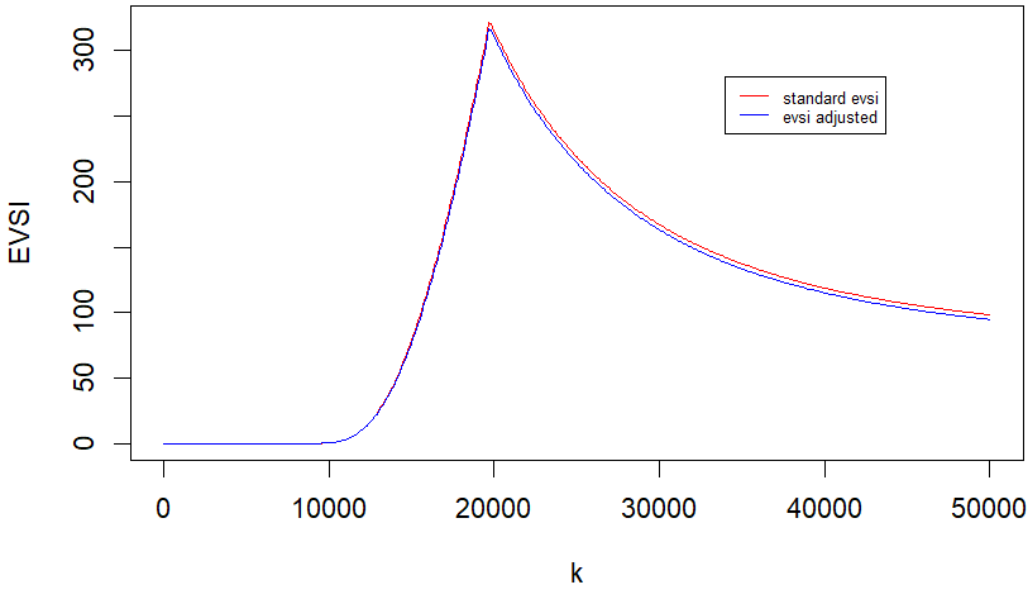


Figure 3.8: EVSI calculated relying on RCT data $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$ in red, EVSI computed on multiple imputed data $(\boldsymbol{Y}^{(i)}_{m}, \boldsymbol{X}, \boldsymbol{Z})$ under MAR assumption with adjusted sample size in blue

Analogously to the previous application, we demonstrated that the original level of EVSI has been achieved by matching the second method $\tilde{T}(\boldsymbol{Y}_m)$ with $T(\boldsymbol{Y}_{RCT})$. Note that the original sample size in this case was $n = 1000$ with a percentage of missingness $p_0 = 0.5$, and the adjusted sample size to match the variance of $T(\boldsymbol{Y}_{RCT})$ has been estimated in $n_{adj} \approx 2450$. As we highlighted in the previous application, in studies affected by missingness, the adjustment researchers make follows this relation $n_{adj} \approx n/(1 - p_0)$, meaning that using a standard approach to sample size calculation in this specific scenario, we would have collected data with a sample size of 2000, with the consequence of not reaching the original level of EVSI. This deviation is way greater in this application with respect to the previous one.

3.6 Conclusions

In this Chapter, we have introduced a methodology to extend the computation of EVSI to more complex data collection mechanisms. In particular, we developed EVSI calculations when the data are affected by missing data, which is a very common case in realistic applications.

When computing EVSI, the results indicate a strong need to account for the additional level of uncertainty induced by primarily two different layers, as it is shown that the additional uncertainty implies a decrease in the EVSI. The first layer that causes this decrease is related to the additional variance introduced by the analysis method we employ to handle missing data, specifically multiple imputation. The second is related to the uncertainty in the missing mechanism, that is, we don't know in advance which observations will be missing. In order to accurately assess the value of information for these types of data and to build a realistic study design, we must take into account primarily these effects.

A methodological solution is also proposed to mitigate the effect of missing data and

to achieve the level of information we would have without missingness. In particular, we proposed a computationally efficient method to determine the number of individuals required to match the information from an RCT with a specific sample size, thereby recovering the EVSI value of data with no induced missingness.

In this context, it would be interesting to explore theoretical properties that could help better define, identify, and quantify the additional sources of uncertainty contributing to the decrease in EVSI. Moreover, these properties would be required to determine a formal technique for quantifying the loss in EVSI due to the different types of missing generating mechanisms (MCAR, MAR, MNAR), or, in other words, the economic value of missing data.

Another interesting direction to explore is related to the case of MNAR missing mechanism, to analyze its effect both when we assume we know the true missing mechanism and when we do not, thereby obtaining a biased estimate of the statistic T, and in general, of the relevant parameters. This is particularly relevant since a biased estimate of the statistic T in the nonparametric method to compute EVSI does not directly introduce an additional layer of uncertainty; however, it still impacts the estimation of EVSI. In this case, additional considerations and investigations must be taken into account.

Finally, different methods, rather than multiple imputation, can be tested to address missing observations. In this way, we can evaluate which method is better at retrieving the initial level of EVSI, thereby defining another way to compare different methods used to deal with missingness.

References

Ades, A., Lu, G., and Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling. *Medical decision making*, 24(2):207–227.

Baio, G. (2013). *Bayesian methods in health economics*, volume 15. CRC Press Boca

Raton (FL).

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365.

Bliss, C. I. (1934). The method of probits. *Science*, 79(2037):38–39.

Fiero, M. H., Huang, S., Oren, E., and Bell, M. L. (2016). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*, 17:1–10.

Gabrio, A., Mason, A. J., and Baio, G. (2017). Handling missing data in within-trial cost-effectiveness analysis: a review with future recommendations. *PharmacoEconomics-open*, 1(2):79–97.

Heath, A., Kunst, N., and Jackson, C. (2024). *Value of information for healthcare decision-making*. CRC Press.

Heath, A., Manolopoulou, I., and Baio, G. (2019). Estimating the expected value of sample information across different sample sizes using moment matching and nonlinear regression. *Medical Decision Making*, 39(4):347–359.

Heath, A., Strong, M., Glynn, D., Kunst, N., Welton, N. J., and Goldhaber-Fiebert, J. D. (2022). Simulating study data to support expected value of sample information calculations: a tutorial. *Medical Decision Making*, 42(2):143–155.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Molenberghs, G. and Kenward, M. (2007). *Missing data in clinical studies*. John Wiley & Sons.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (2018). Multiple imputation. In *Flexible imputation of missing data, second edition*, pages 29–62. Chapman and Hall/CRC.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Strong, M., Oakley, J. E., Brennan, A., and Breeze, P. (2015). Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Medical Decision Making*, 35(5):570–583.

Chapter 4

Observational Expected value of Sample Information

4.1 Introduction

As mentioned in Chapter 3, the EVSI methodology (Ades et al. (2004)) has been applied primarily to data from RCTs or other simplified data collection mechanisms (Heath et al. (2022)). For this reason, in Chapter §3, we introduced a methodology to perform EVSI when planning to collect more realistic data, that is, data with missing values(Rubin (1976)). However, we already highlighted that RCTs can suffer from high costs and difficulties in feasibility in different contexts. Moreover, there is a range of alternative study types that we would like to use in our EVSI calculations. Therefore, in this chapter, we define a novel methodology for applying EVSI when planning to collect observational data (Rosenbaum et al. (2010)).

Understanding how to perform EVSI in face of different types of observational data, with varying associated biases, can be particularly useful, as it would allow us to utilize such a powerful instrument in contexts where we are not as concerned about the accessibility and costs of data (Benson and Hartz (2017)). Moreover, as with the computation on RCTs, understanding how to apply EVSI in this novel context would be highly beneficial in designing experiments when planning to collect observational data (research prioritization) (Claxton and Sculpher (2006)).

The methodology for computing EVSI on observational data follows a very similar structure to that of the previous chapter, but with a significant difference. In this context, to define the data-generating process, we will apply inverse target trial emulation (ITTE), a methodology defined in Chapter 2, to generate observational data from existing or synthetic RCT data.

Once data are generated, we compare the results obtained by performing EVSI on the original RCT with those obtained from the observational data using ITTE. This will help us understand the potential drawbacks of using EVSI with these types of data and how to address them. Notice that, to apply EVSI to observational data, we will proceed as we

usually do when analyzing real-world data, that is, we address the bias within the data using different techniques of Target Trials Emulation (TTE), and we compute EVSI on the final data. Even if the methodology that we will define can potentially include any type of observational data, in this work, we focus on a particular type, which is related to imbalance in populations originating from confounding §1.2.1.1. In this context, the methods that one can apply to properly analyze the observational data include matching, inverse probability weighting, regression methods, and g-methods (Rubin (1974), Rubin (1973), Imbens and Rubin (2015)). In this work, we will apply inverse probability weighting (IPW) (Chesnaye et al. (2022)). For a broader application of the other methods to address different ITTE techniques, see Chapter 1.

The chapter proceeds as follows: first, we recall ITTE's description of how observational data can be generated in the presence of population imbalance and how this bias can be mitigated using various techniques in different contexts. Second, we introduce a general setting with an underlying health-economics model and a Bayesian data-generating model to compute EVSI. Third, we apply this methodology in two different health-economic settings. In the first, we will assume a conjugate Normal Normal model for the data-generating process. In contrast, the second scenario refers to a realistic setting where the underlying economic model is the Markov model introduced in §1.1.2. We represent the EVSI of the starting RCT, and we compute and represent the EVSI of the final data, i.e., the ones obtained by adjusting for biases in the simulated observational data. Finally, we discuss the results.

4.2 Inverse Target Trial emulation and unbalanced populations

4.2.1 Inverse target trial emulation

Observational data often exhibit different types of bias, such as *confounding* (Walker (1996), Assimon (2021)), *selection bias* (Hernán et al. (2004), Hegedus and Moody (2010)), and *information bias* (Althubaiti (2016), Tripepi et al. (2010)). These three broad categories encompass different types of bias (Sackett (1979), Grimes and Schulz (2002)). For this reason, it is essential to develop a methodology for properly simulating and analyzing this type of data.

In Chapter 2, we introduced *Inverse target trial emulation*, a methodology for generating observational data from preliminary RCT data. In particular, we focused our attention on reproducing two relevant types of confounding, selection, and information bias or effect of those: *time zero bias* (Suissa (2008)) and *unbalanced population* (Rosenbaum et al. (2010)). In the standard methodology, we assumed having access to a preliminary RCT; however, in general, we can also simulate a synthetic RCT and use it for implementing the ITTE methodology.

This section describes how ITTE works when generating observational data with different features (bias). In particular, we focus on the situations in which data exhibit imbalances in populations, a very common situation when data are affected by different forms of confounding and/or selection bias.

First, we build synthetic RCT data applying the same nested model of §3.2.1.1. Note that if we have access to real RCT data, the nested model can be used to infer the relevant parameters directly from the data (as in the original ITTE). In this case, the RCT should satisfy some basic requirements to ensure that these parameters are informative for the simulation study. In particular, the RCT should present population, intervention,

comparison and measured outcomes comparable to the target population of interest. In this case we would be able to infer the distribution of covariates and the relationships between covariates and outcomes from the RCT. If unmeasured but relevant confounders are suspected, they could be incorporated into the simulation by specifying plausible distributions informed by expert opinion or literature evidence, and their impact explored through sensitivity analyses. Otherwise, the standard no unmeasured confounding (or conditional exchangeability) assumption must be invoked. Conversely, if we do not have an initial RCT that satisfies these conditions, relevant parameters can be informed using the existing literature. For example, published epidemiological studies, observational cohorts, meta-analyses or different literature evidence. Similarly, treatment effect estimates from clinical trials in related populations can be used to define plausible ranges or prior distributions for the relevant model components. In practice, evidence from multiple sources will often need to be combined, ensuring internal coherence between covariate distributions, outcome models, and expected treatment effects.

Let us consider a general RCT with a population of N individuals. For each individual, the randomized data contain S covariates $\{X_1, X_2, .., X_S\}$, the exposure variable Z and the outcome variable Y.

As we describe in §2.2.2.1 and recall in the previous chapter, we first factorize the joint distribution of the covariates and the outcome variable in a product of sequential conditional distributions. For example, the k-th component is $X_k|\boldsymbol{X}_{1,2,...,k-1} \sim F_k(\boldsymbol{\Theta}_{k-1}(\boldsymbol{X}_{1,2,...,k-1}))$ and the the conditional distribution of the outcome $Y|\boldsymbol{X}, Z \sim F(\boldsymbol{\Theta}(\boldsymbol{X}, Z))$, for details see the previous chapters. Note that in the EVSI computation outcome depends on the PA simulations of the relevant parameters, as detailed in the next section.

Once we have generated RCT data $(\boldsymbol{Y}_{\text{RCT}}, \boldsymbol{X}, \boldsymbol{Z})$, we can then decide which type of observational data we want to emulate by reproducing different types of bias that can occur in a realistic scenario, we now delve into one specific type of bias: the presence of

unbalanced populations.

4.2.2 Unbalanced populations

To generate observational data with unbalanced populations from an RCT, we can proceed with different techniques as described in §2.3.1. We recall that each method operates under different assumptions about how imbalance arises within the observational data we aim to reproduce, i.e., which types of imbalance we are expected to observe in the data we collect.

In this setting, we assume that we know which characteristics of the patients (covariates) generate an imbalance between the treated and control populations. Suppose, for example, that in our setting, older individuals are treated more than younger ones, therefore unbalancing the populations with respect to patients' ages. In §2.3.1.2, we defined two different methods to induce population imbalance in RCT data in this specific scenario: *Decoupling Mahalanobis* and *Decoupling covariate*.

We focus our attention on the *Decoupling covariate*, which explicitly specifies which covariate(s) induces differences in the control and treated groups. In this context, we want to introduce a stochastic version of the above-mentioned method, as in the EVSI computation process, we usually do not have prior knowledge of the precise degree of overlap between populations or the precise manner in which the selected covariates generate imbalance.

The stochastic version, namely *Stochastic decoupling covariate*, described in Table 4.1, works by assigning a higher probability of being discarded from the dataset to people with a specific covariate profile in the control group, and a lower probability to people with the same covariate profile but in the treatment group.

Note again that it is essential to consider a stochastic version of the method in §2.3.1.2, as in the EVSI calculation, we must consider all the different sources of uncertainty and variability in the data-generating process. This is because, in realistic scenarios, we usually

Stochastic decoupling covariate

1. Start with a synthesized RCT D_{RCT} and select a particular covariate $\boldsymbol{X}^{(i)}$.

2. Define the probability $p(x^{(i)})$ of an individual with covariate profile $x^{(i)}$ to be discarded so that

$$p(x_s^{(i)}) > p(x_t^{(i)}) \quad if \quad x_s^{(i)} > x_t^{(i)} \quad and \quad Z_s = Z_t = 1$$

$$p(x_s^{(i)}) > p(x_t^{(i)}) \quad if \quad x_s^{(i)} < x_t^{(i)} \quad and \quad Z_s = Z_t = 0$$

   Note that, without loss of generality, we assume that the higher values of the covariates lead to higher probability of inclusion in the dataset. Moreover,

$$p(max(X_s^{(i)})) = p(min(X_t^{(i)})) = 1 \quad for \quad Z_s = 1, Z_t = 0$$

   Represents the probability of an individual with covariate $X^{(i)} = x^{(i)}$ to be discarded from the data;

3. For each individual sample $\pi_j = Bern(p(x_j^{(i)}))$, the missing indicator ($\pi_j = 1$ if individual j misses and $\pi_j = 0$ otherwise.

4. Select individuals $\{I_1, .., I_k\}$ from D_{RCT} such that $\pi_{I_i} = 1$ (missing indicator for individual I_i) and $Z_{I_i} = 1$ for $i = 1, .., k$, and h individuals $\{J_1, .., J_h\}$ from D_{RCT} such that $\pi_{J_i} = 1$ and $Z_{J_i} = 0$ for $i = 1, .., h$.

5. Remove the selected individuals from D_{RCT}.

6. Repeat until the prespecified number N_{dis} of individuals to discard from each population is reached and D_{obs} is obtained.

Table 4.1: Stochastic decoupling covariate method

do not know precisely in advance how the non-overlap between populations originates.

Related to this, note that with this technique, we assume we know which covariates produce the non-overlap between the two populations and have an approximate idea of how they distribute among the two groups. We do not assume to know precisely which is the conditional exposure distribution given the observed confounders. Again, notice that in §2.3.1, different techniques, working under various degrees of knowledge, are described, but for the sake of this article, we limit ourselves to the one just described.

As we introduced in §1.2.1.1 and mentioned above, various methods can be applied to adjust for population imbalance. In this work, we will employ inverse probability weighting.

4.3 VoI analysis with observational data

4.3.1 Standard context

We recall the decision-analytic framework introduced in §1.1.1 and reiterated in §3.2.

Let's consider an underlying health economic model governed by parameters $\boldsymbol{\theta}$, which we assume to have a joint distribution p that represents the uncertainty associated with the set of parameters.

In this context, we assume that we need to compare a set of D different decisions, $d = 1, \ldots, D$, each of which is associated with a specific net benefit $\text{NB}_d(\boldsymbol{\theta})$. Recall that the definition of the EVSI, introduced in §1.3.1, is

$$\text{EVSI} = \mathbb{E}_{\boldsymbol{Y}}\left[\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left(\text{NB}_d(\boldsymbol{\theta})\right)\right] - \max_d \mathbb{E}_{\boldsymbol{\theta}}\left(\text{NB}_d(\boldsymbol{\theta})\right)$$

Where $\boldsymbol{Y}$ is the data we plan to collect to inform our decision model.

As we mentioned in §1.3.2, there are several ways to compute EVSI. In this work, we consider the nonparametric regression-based method (Strong et al. (2015)). In this

approach, multiple nonparametric regression equations (one for each possible treatment) are fitted to compute the preposterior mean:

$$\mathrm{NB}_d(\boldsymbol{\theta}) = g_d(T(\boldsymbol{Y})) + \epsilon$$

where $\boldsymbol{\theta}$ is a sample of the joint distribution of the input parameters obtained by performing a probabilistic analysis (PA), and T is some low-dimensional summary statistic of the data as deepened in §1.3.2.

As mentioned above, previous examples of calculating EVSI have only generated $\boldsymbol{Y}$ from simple data collection schemes, which can be unrealistic in concrete scenarios. In Chapter 3, we introduced a novel methodology to compute EVSI with data affected by missingness. In this work, we aim to introduce a methodology that enables us to compute the value of collecting more data from non-experimental studies, allowing us to apply a powerful instrument such as EVSI in contexts where we do not have to worry about data availability and cost.

4.3.2 EVSI with observational data

When we deal with observational data, the above process must change. If we plan to collect observational data, then we must modify the standard process above by sampling observational data ($\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs}$) instead of RCTs. This step will be done using ITTE as introduced in the previous section.

Once observational data are sampled, we have successfully reproduced the realistic situation in which a researcher is supposed to collect additional observational data, and they compute the value of information for that data.

In a realistic scenario, before conducting the analysis and economic evaluation, one must address the bias within the data using various techniques, as mentioned, depending on the type of bias encountered. Similarly, when using non-parametric regression methods

to compute EVSI, we cannot directly use $T(\boldsymbol{Y}_{obs})$ in our estimates. Therefore, we must first manipulate our data to find an unbiased estimate of $T(\boldsymbol{Y}_{\text{RCT}})$. As said, we will induce imbalance in populations and apply IPW to obtain an unbiased estimate of $T(\boldsymbol{Y}_{\text{RCT}})$.

We denote as $(\boldsymbol{Y}_{RCT}, \boldsymbol{X}, \boldsymbol{Z})$ the starting RCT, simulated at the beginning of the process; as $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs})$, the observational data obtained by ITTE, and as $\tilde{T}(\boldsymbol{Y}_{obs})$ the unbiased estimate of $T(\boldsymbol{Y}_{RCT})$.

The methodology we introduce for computing EVSI with observational data has the same structure as the one introduced in Chapter 3. Even if the underlying logic is the same, we have to change the manipulation of the simulated RCT data to generate, this time, observational data. We briefly outline the different steps of the novel methodology, focusing on the steps that differ from the methodology defined in §3.4.1.

1. Generate covariates $\boldsymbol{X}$ and the exposure variables $\boldsymbol{Z}$ for N individuals $j = 1, .., n$.

2. Sample K parameters $\boldsymbol{\theta}_i$, $i = 1, \ldots, K$ inside the PA simulations.

3. Develop the individual-level parameters $\tilde{\boldsymbol{\theta}}_i$ required to simulate the outcomes, inducing dependence of the relevant parameters with covariates as exposed in §3.4.1.

4. Sample new data (the outcome or another relevant measure) governed by the relevant parameters, as in the standard approach, coming from an RCT

$$\boldsymbol{Y}_{RCT}^{(i)} \sim F(\tilde{\boldsymbol{\theta}}_i)$$

   and obtain K starting RCTs $(\boldsymbol{Y}_{RCT}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ and $\boldsymbol{Z}$ have been fixed at the beginning of the process.

5. Depending on which type of bias, i.e. which type of observational data, we are supposed to observe, we transform the RCTs obtained at the previous point in observational data by ITTE, as shown in §3.2.

6. Once we have $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs})^{(i)}$ we can apply different techniques (IPW, Mahalanobis matching,...), depending on the bias we are facing, to adjust data and compute $\tilde{T}(\boldsymbol{Y}_{obs})$, an unbiased estimator of $T(\boldsymbol{Y}_{\text{RCT}})$.

7. Estimate preposterior mean with nonparametric regression methods as and obtain an estimate of $EVSI|\boldsymbol{X}$.

8. Finally we integrate out the dependence on $\boldsymbol{X}$ and obtain the final estimate of the EVSI.

The relevant differences, with respect to EVSI and missing data, are those in points 5 and 6.

In the previous approach, we chose the missing mechanism and then applied multiple imputation to estimate the missing observations. In this context, we apply Inverse Target Trial Emulation to completely transform the data structure, thereby reproducing all the different types of bias that may be encountered when dealing with observational data. With this methodology, we can induce all the relevant types of bias described in Chapter 2. Moreover, whenever ITTE is generalized to include other types of bias, we can potentially incorporate them within this methodology.

As we have already specified, in this work, we will focus on a particular type of bias, namely *Imbalance in populations*, and apply the methodology above in the same different settings as those in Chapter 2, that is, the Normal conjugate model and the Chemotherapy Markov model.

4.4 Application

We apply the novel methodology to compute EVSI with observational data in a Normal-Normal conjugate model and Chemotherapy treatment model.

4.4.1 Observational EVSI Normal conjugate model

We do not explicitly recall the Normal-Normal conjugate model; for the extensive definition, see §3.4.1.

To describe how the computation of EVSI in this context, we suppose again that we will collect data on a set of $n = 100$ individuals, with different ages *age* and levels of comorbidity *com*, and measure the effectiveness Y of a specific therapy in these individuals.

Following the same scheme of the previous section, we list the explicit steps, noting that steps 1-4 are the same as the ones in §3.4.1.

1. We simulate two independent covariates for each individual j: *age, com*, and we assume that the first half of the individuals are treated and the other half are not. We set $n_0 = 50$ and $n_1 = 50$, respectively, the number of people in the control and treatment group, and $n = n_1 + n_2$.

$$age_j \overset{\text{iid}}{\sim} \mathcal{N}(57, 100) \quad j = 1, ..., n$$

$$com_j \overset{\text{iid}}{\sim} Bin(10, 0.5) \quad j = 1, ..., n$$

2. Then, we sample $K = 5000$ PA simulations of relevant parameters $\theta^{(i)}_{Z=1}$ and $\theta^{(i)}_{Z=0}$:

$$\theta^{(i)}_{Z=0} = \theta^{(i)}|Z = 0 \sim N(\theta_0, \sigma_0^2) \quad i = 1, .., K$$

$$\theta^{(i)}_{Z=1} = \theta^{(i)}|Z = 1 \sim N(\theta_1, \sigma_1^2) \quad i = 1, .., K$$

   with $\theta_0 = 20$, $\theta_1 = 40$, $\sigma_1 = 8$ and $\sigma_0 = 6$.

3. For each simulated individual j, we compute $\tilde{\theta}_i$ (the mean of the outcome distribu-

tion, given the patient's i characteristics) as follows:

$$\tilde{\theta}_j^{(i)} = \theta_{Z_j}^{(i)} + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) \quad j = 1, ..., n$$

where $\theta^{(i)}$ is the i-th sample of θ inside PA simulation and $\alpha_1 = 5, \alpha_2 = 8$

4. We sample the outcome for individual j as:

$$Y_j^{(i)} \overset{\text{iid}}{\sim} N(\tilde{\theta}_j^{(i)}, \sigma_Y^2)$$

and obtain $K = 5000$ starting RCTs $(\boldsymbol{Y}_{\text{RCT}}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ were fixed at the beginning of the process.

5. We decide with which method we introduce bias and separate populations, as we suggested in §4.2.2 we proceed with the **stochastic decoupling covariate** method. In this case, we decide to remove m individuals based on covariate age. Following the technique presented in §4.2.2, we will set $p(age_s) \propto age_s/max(age_s)$ when $Z_s = 1$ and $p(age_t) \propto min(age_t)/age_t$ when $Z_t = 1$.

6. Once we obtain the simulated observational data $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs})^{(i)}$, for $i = 1, .., K$, we perform inverse probability weighting to each of the K dataset to obtain a super-population balanced by weighting observations with weights $(\gamma_1, ..., \gamma_n)$. In this case, we get the unbiased estimator of $T(\boldsymbol{Y}_{\text{RCT}})$, that is, $\tilde{T}(\boldsymbol{Y}_{obs})$, by applying weights $\boldsymbol{\gamma}$ to $T(\boldsymbol{Y}_{obs})$.

7. We estimate the preposterior mean from the following regression with Generalized additive model (GAM):

$$\text{NB}(\boldsymbol{\theta}_i) = g(\tilde{T}(\boldsymbol{Y}_{obs})) + \epsilon \tag{4.1}$$

and obtain an estimate of $EVSI|\boldsymbol{X}$;

8. Finally, we integrate out the dependence on $\boldsymbol{X}$ and Z and obtain the final estimate of the EVSI as:

$$\text{EVSI} = \frac{1}{N}\sum_{i=1}^{N} EVSI|\boldsymbol{X}_i$$

This is achieved by computing EVSI$|X$ N times, where a different covariate profile is sampled each time, and then averaging over the obtained EVSI curves. In this and the following application, we set $N = 10$ as in the previous chapter.

4.4.2 Chemotherapy treatment model

Again, for the formal introduction of the Chemotherapy treatment model we refer to the previous chapters, see §1.1.2.

In this case, we suppose that we will collect data on a set of $n = 1000$ individuals, with different *age* and levels of comorbidity *com*, and observe the outcome Y, which represents the occurrence of a side effect. Following the methodology introduced in §4.3.2 we list the relevant steps, again noticing that steps 1-4 are the same as §3.4.2:

1. We simulate two independent covariates for each individual j: *age, com*, and we assume that the first half of the individuals are treated and the other half are not. We set $n_0 = 500$ and $n_1 = 500$, respectively, the number of people in the control and treatment group, and $n = n_1 + n_2$.

$$age_j \overset{\text{iid}}{\sim} \mathcal{N}(57, 64) \quad j = 1, ..., n$$

$$com_j \overset{\text{iid}}{\sim} Bin(10, 0.5) \quad j = 1, ..., n$$

2. then, we sample $K = 5000$ PA simulations with relevant parameters sampled from

$$log(\rho^{(i)}) \sim N(log(0.54), 0.3^2) \quad i = 1, .., K$$

$$s_0^{(i)} \sim Beta(1, 1) \quad i = 1, .., K$$

$$s_1^{(i)} = 1 - \frac{1 - s_0^{(i)}}{s_0^{(i)}(\rho^{(i)} - 1) + 1} \quad i = 1, .., K$$

and all the other relevant parameters listed in §1.1.2. ρ is the odds ratio of a side effect occurrence, π_0 is the prior probability of a side effect occurrence in the control group, while π_1 is the analogous for the treated group. For a complete list of the parameters and their distributions, we refer to (Heath et al. (2024)).

3. For each simulated individual j, we compute $\tilde{s}_j^{(i)}$ (the mean of the outcome distribution, that is the probability for individual j of having a side effect, given the patient's j characteristics) as follows:

$$\tilde{s}_{0,j}^{(i)} = \Phi(\Phi^{-1}(s_0^{(i)}) + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) + \alpha_{0,0}) \quad j \text{ s.t. } Z_j = 0$$

where $s_0^{(i)}$ is the i-th sample of PA simulation (for control group), $\alpha_1 = 0.05$, $\alpha_2 = 0.3$, Φ is the CDF of a standard Gaussian and $\alpha_{0,0}$ is such that:

$$\sum_{j=1}^{n} \Phi(\Phi^{-1}(s_0^{(i)}) + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) + \alpha_0) - s_0^{(i)} = 0$$

so that the individual parameters preserve the overall mean $s_0^{(i)}$. In an analogous way we define:

$$\tilde{s}_{1,j}^{(i)} = \Phi(\Phi^{-1}(s_1^{(i)}) + \alpha_1(age_j - \overline{age}) + \alpha_2(com_j - \overline{com}) + \alpha_{0,1}) \quad j \text{ s.t. } Z_j = 1$$

4. we sample the outcome for individual j as:

$$Y_j^{(i)} \overset{\text{iid}}{\sim} Bern(\tilde{s}_j^{(i)})$$

and obtain $K = 5000$ starting RCTs $(\boldsymbol{Y}_{\text{RCT}}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ for $i = 1, .., K$. Note again that $\boldsymbol{X}$ were fixed at the beginning of the process.

5. as suggested in §4.2.2, to introduce bias and separate populations, we proceed with the **stochastic decoupling covariate** method. Again, we decide to remove m individuals based on the covariate age and set $p(age_s) \propto age_s/max(age_s)$ when $Z_s = 1$ and $p(age_t) \propto min(age_t)/age_t$ when $Z_t = 1$.

6. once we obtain the simulated observational data $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs})^{(i)}$, for $i = 1, .., K$, we perform inverse probability weighting to obtain a superpopulation balanced by weighting observations with weights $(\gamma_1, ..., \gamma_n)$. In this case, we get the unbiased estimator of $T(\boldsymbol{Y}_{\text{RCT}})$, that is $\tilde{T}(\mathbf{Y}_{obs})$, by applying weights $\boldsymbol{\gamma}$ to $T(\boldsymbol{Y}_{obs})$. Note that in this case, T is no longer the sample mean, as in the Normal conjugate model, but the odds ratio;

7. finally, we integrate out the dependence on $\boldsymbol{X}$ and Z and obtain the final estimate of the EVSI as:

$$\text{EVSI} = \frac{1}{N} \sum_{i=1}^{N} \text{EVSI}|\boldsymbol{X}_i$$

4.4.3 Comparison and second moment matching

In the next section, we will apply the novel methodology to Normal-Normal and Chemotherapy models. We will compare the original EVSI, that is the one obtained by using the RCT data $(\boldsymbol{Y}_{\text{RCT}}^{(i)}, \boldsymbol{X}, \boldsymbol{Z})$ sampled in point 4, with the EVSI computed by applying IPW on observational data $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs})$ to obtain the unbiased estimator $\tilde{T}(\boldsymbol{Y}_{obs})$.

Note that, in the *stochastic decoupling covariate* method, we discard N_{dis} individuals from each population to generate unbalanced populations. For this reason, we will compare the original EVSI with the EVSI computed on an initial sample size of $\tilde{n} = n+2N_{dis}$, that is, by adjusting the process so that we obtain n individuals after the decoupling method, therefore reaching the original sample size and ensuring fair comparisons.

Moreover, we recall from §2.5 that even though the method used there was slightly different, discarding more individuals in *stochastic decoupling covariate* consistently leads to a reduced degree of overlap between the populations, or increases the degree of non-overlap (measured by population Mahalanobis distance). In the next section, we will also compare the EVSI computed with observational data obtained by discarding two different numbers of individuals in *stochastic decoupling covariate*, obtaining different degrees of non-overlap.

Finally, as in the previous section, we estimate the sample size we need to reach the initial level of EVSI by matching the second moment of $\tilde{T}(\boldsymbol{Y}_{obs})$ to the one of the original $T(\boldsymbol{Y}_{\text{RCT}})$ for an increasing sample size of $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs})$. Numerical results show that this time, the computation of variance appears to be more sensitive to simulated confounders, suggesting that one must exercise greater caution when drawing conclusions.

4.5 Results

4.5.1 Normal conjugate model

We start by comparing the EVSI curves of the following quantities:

1. Original EVSI (equal to EVSI computed on RCT data with no decoupling);
2. EVSI computed on data obtained through inverse probability weighting.

Since we discarded $N_{dis} = 20$ individuals from each population to generate unbalanced populations, we will compare the original EVSI with the EVSI computed with

$\tilde{n} = 100 + 2N_{dis} = 140$, that is, by adjusting the process to obtain 100 individuals after the decoupling method reaching the target sample size.

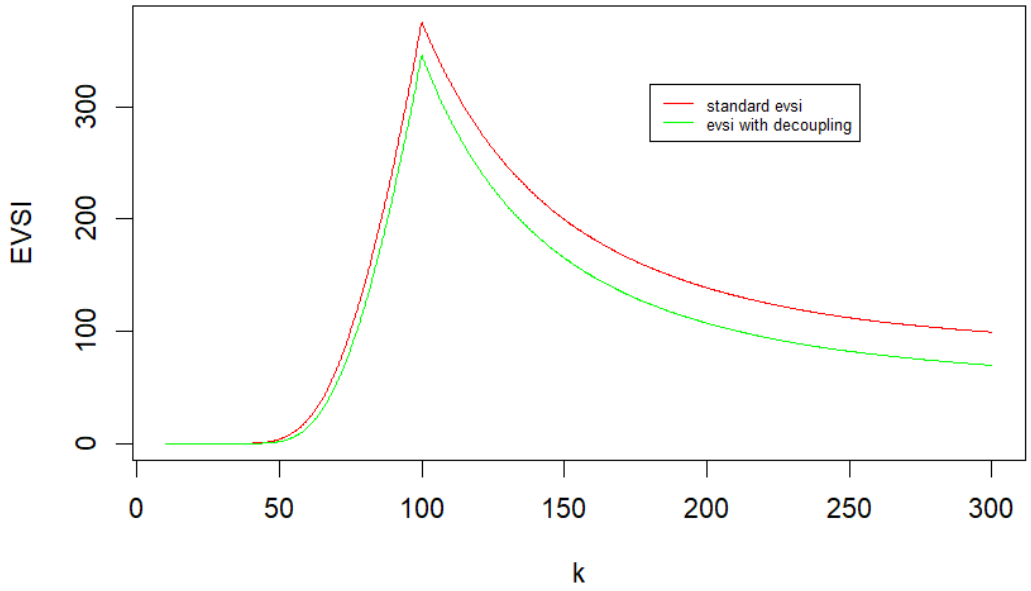


Figure 4.1: EVSI on simulated RCTs $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \mathbf{Z})$ for $i = 1, .., K$ in red and EVSI computed on data obtained through stochastic decoupling covariate $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \mathbf{Z}_{obs})^{(i)}$ with $N_{dis} = 20$ in green. Sample size $n = 100$.

The results show a decrease in the EVSI. Similar to the case of the EVSI on RCT with missing values, this decrease in the economic value of information can be explained by the additional layers of variability induced by both the Stochastic decoupling method in §4.2.2 and the use of IPW to obtain unbiased estimates.

To determine the sample size necessary to reach the previous level of information we proceed as in the last chapter, that is, we match the second moment of $\tilde{T}(\boldsymbol{Y}_{obs})$ compared to the one of the original $T(\boldsymbol{Y}_{RCT})$ (the one with no decoupling).

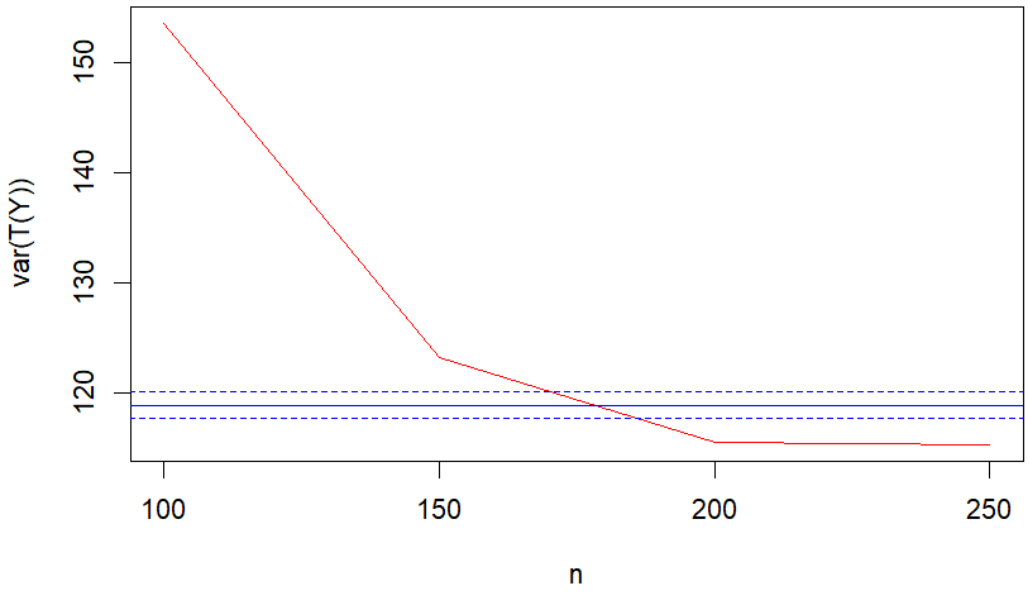


Figure 4.2: Comparison between $\text{var}(\tilde{T}(\boldsymbol{Y}_{obs}))$ for increasing sample size n and $\text{var}(T(\boldsymbol{Y}_{RCT}))$.

To match the second moment, it is necessary to collect $n = 180$ individuals, which means we need 180 individuals to match the information from an RCT with 100 patients. Then, we compute the EVSI with the adjusted sample size to demonstrate that the previous level of value of information has been reached.

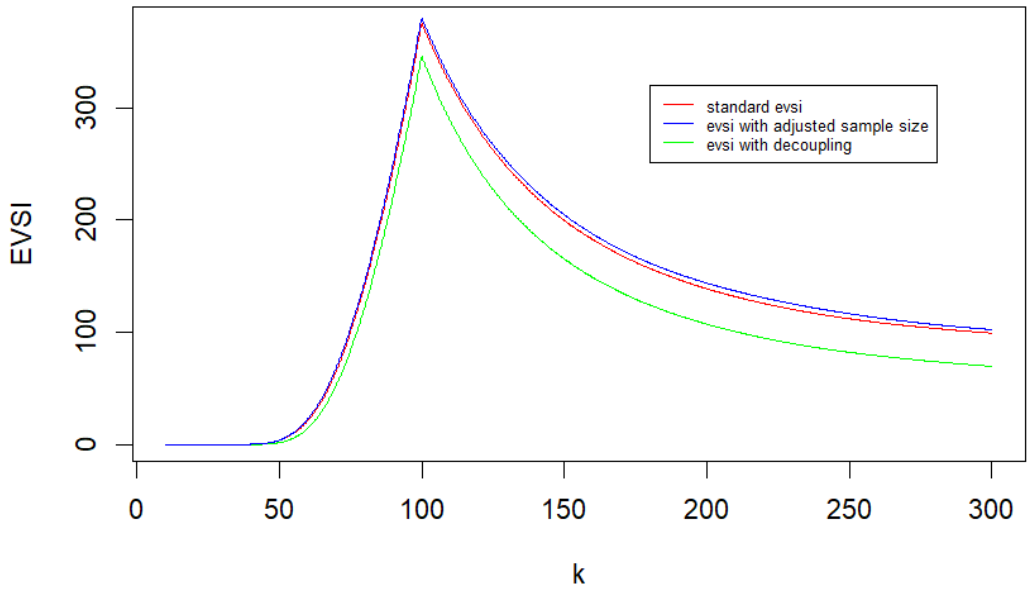


Figure 4.3: EVSI calculated relying on RCT data $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \mathbf{Z})$ for $i = 1, .., K$ in red, EVSI computed on data obtained through stochastic decoupling covariate $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \mathbf{Z}_{obs})^{(i)}$ with adjusted sample size $\tilde{n} = n + 2N_{dis}$ in green and EVSI computed on data obtained through stochastic decoupling covariate with $N_{dis} = 20$ in green.

4.5.2 Chemotherapy model

In the Chemotherapy model, we assume we would collect $n = 1000$ individuals, compared to the 100 of the Normal conjugate model. This is because, now that we are dealing with observational data, we can expect a larger sample size due to higher availability and feasibility.

Again, we compare the original EVSI (without decoupling) with EVSI on observational data with unbalanced populations. This time, we discard $N_{dis} = 1000$ individuals from each population to generate unbalanced populations; therefore, we compare the original EVSI with the EVSI computed with $\tilde{n} = 1000 + 2N_{dis} = 3000$, that is, by adjusting the process to obtain 1000 individuals after the decoupling method reaches the original sample size.

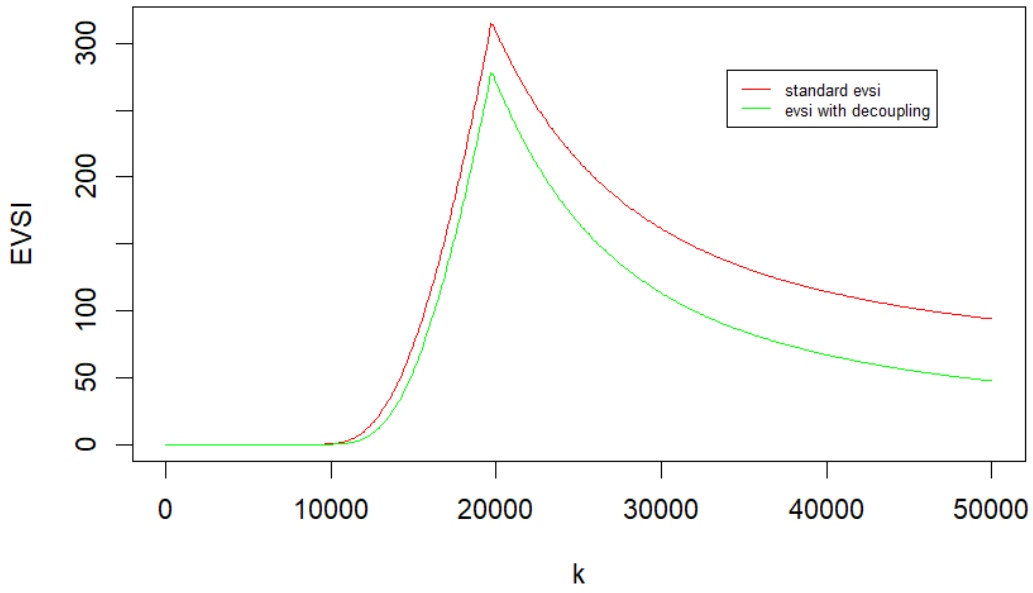


Figure 4.4: EVSI on simulated RCTs $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \mathbf{Z})$ for $i = 1, .., K$ in red and EVSI computed on data obtained through stochastic decoupling covariate $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \mathbf{Z}_{obs})^{(i)}$ with $N_{dis} = 500$ in green. Sample size $n = 1000$.

As we have shown, the decrease in EVSI is also observed for larger sample sizes. This comes from the same level of additional variability we referred above.

Note that, since it is known that IPW may cause weight inflation when populations are very different, in our implementation, every time large weights are generated, the simulation is repeated. In all the $K = 5000$ simulations, on average, less than 5% present this well-known problem of IPW.

In §2.5, we demonstrated that the greater the number of individuals discarded, the greater the non-overlap (measured by population Mahalanobis distance) increases. Even if, in the previous case, we only discarded individuals without the need to reach the same sample size after leaving out individuals, we still observe a similar behavior.

In the following simulation, we compare the original EVSI with the EVSI obtained by first discarding $N_{dis} = 500$ individuals per population, and then again $N_{dis} = 1000$. For

each simulation, we also measure the mean non-overlapping degree among the $K = 5000$ simulations. In the case of $N_{dis} = 500$, the degree of nonoverlap is $\Delta_{500} = 1.149$, while for $N_{dis} = 1000$, $\Delta_{1000} = 1.21$

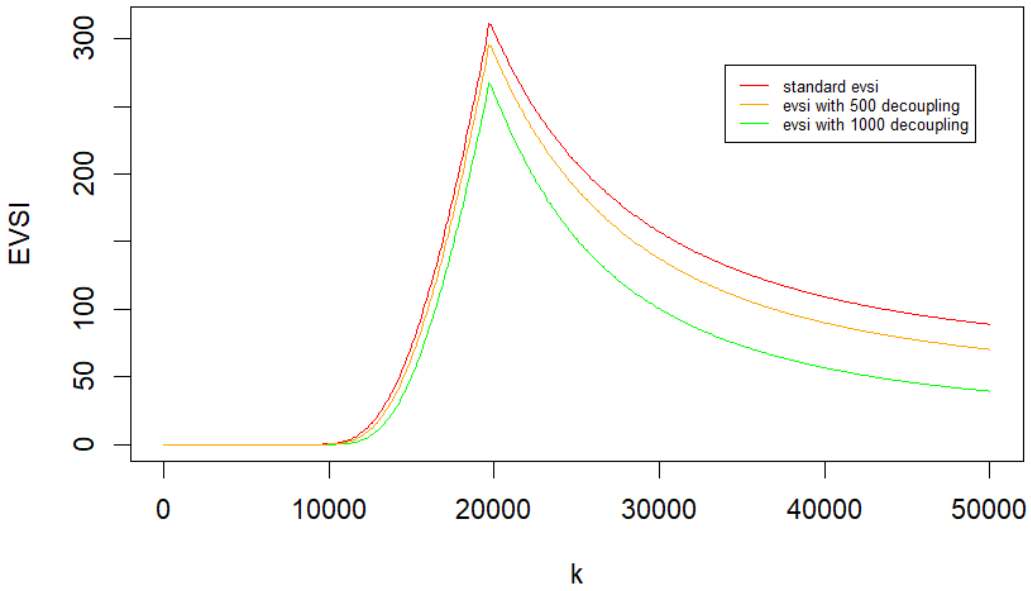
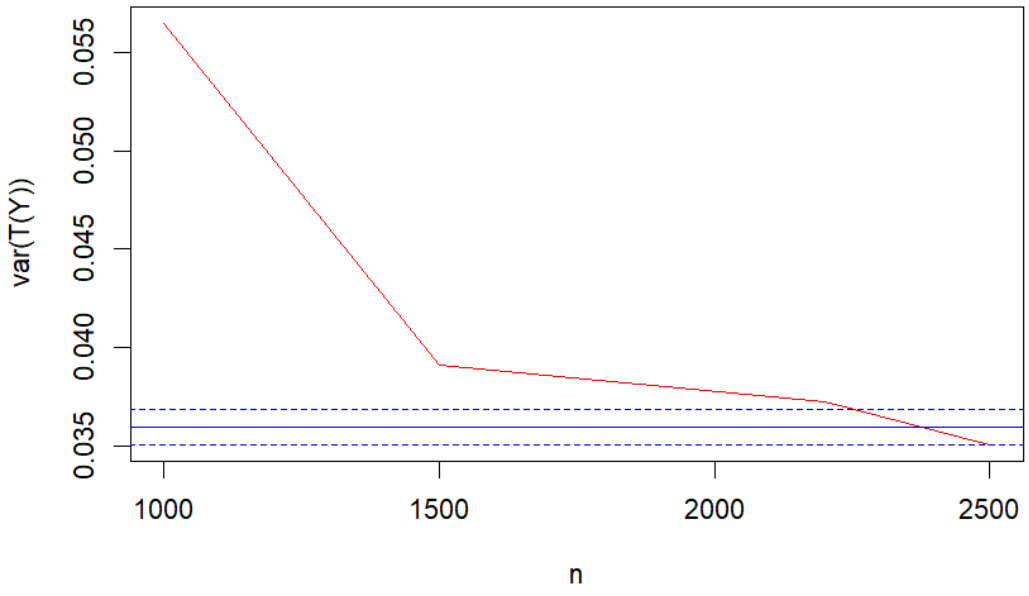


Figure 4.5: EVSI on simulated RCTs $(\boldsymbol{Y}^{(i)}_{\text{RCT}}, \boldsymbol{X}, \mathbf{Z})$ for $i = 1, .., K$ in red, EVSI computed on data obtained through stochastic decoupling covariate $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \mathbf{Z}_{obs})^{(i)}$ with $N_{dis} = 500$ in orange and EVSI computed on data obtained through stochastic decoupling covariate $(\boldsymbol{Y}_{obs}, \boldsymbol{X}_{obs}, \mathbf{Z}_{obs})^{(i)}$ with $N_{dis} = 1000$ in green.

As we observe, the fewer individuals are discarded, the lower the loss in value of information; or, in other terms, the lower the degree of nonoverlap, the lower the loss in EVSI.

Finally, we apply the same method above to estimate the necessary sample size to reach the original level of EVSI.

In this case, we estimate $n \approx 2400$ and we compute the EVSI for this sample size, showing that the original level of value of information is reached.



.

Figure 4.6: Comparison between $\text{var}(\tilde{T}(\boldsymbol{Y}_{obs}))$ for increasing sample size n and $var(T(\boldsymbol{Y}_{RCT}))$

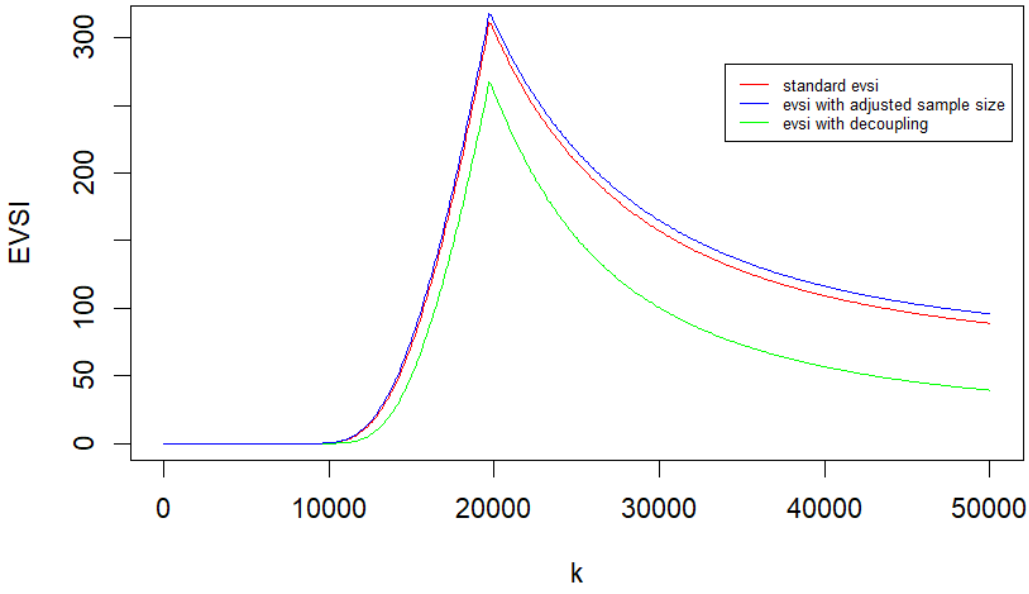


Figure 4.7: EVSI calculated relying on RCT data $(\boldsymbol{Y}_{\text{RCT}}, \boldsymbol{X}, \mathbf{Z})^{(i)}$ for $i = 1, .., K$ in red, EVSI computed on data obtained through stochastic decoupling covariate $(\boldsymbol{Y}_{obs}^{(i)}, \boldsymbol{X}_{obs}, \mathbf{Z}_{obs})$ with adjusted sample size $\tilde{n} = n + 2N_{dis}$ in blue and EVSI computed on data obtained through stochastic decoupling covariate with $N_{dis} = 1000$ in green.

4.6 Conclusions

In this chapter, we introduced a methodology for computing EVSI when we one plans to collect observational data. To achieve this, we combined the results of Chapters 2 and 3. In particular, we apply the general methodology to compute EVSI in non-standard contexts introduced in Chapter 3, with the ITTE methodology to generate observational data introduced in Chapter 2. Even if this is a preliminary yet promising result, combining the high availability and feasibility of observational data with such a powerful instrument for decision-making and research prioritization, as EVSI does, opens up different potential applications in various real-life scenarios in HTA.

Given these promising results, it would be valuable to explore theoretical properties that could help better define, identify, and quantify the additional layers of uncertainty induced by ITTE within the data-generating mechanism, which explains the decrease in the EVSI. To do this, one could start with Bayesian conjugate models, along with more standard methods of inducing bias, and then move to the analysis and estimation of the different layers in more complex scenarios.

Another interesting direction to explore is related to applying this methodology to observational data with different sources of bias from the one considered in this chapter, in order to extend its usage and cover a larger variety of realistic scenarios. Moreover, for each of the different types of observational data, we can evaluate the ability of different methods to properly adjust bias and maintain the original level of EVSI, defining another criterion to determine the best method to adopt to solve bias in different specific scenarios. Also in this case, it would be interesting to define different properties and justify the corresponding results theoretically.

Finally, since numerical results suggest that this methodology appears to be highly sensitive to the confounders we simulate in the original RCT data, we will also further investigate this property to understand how to properly account for this behavior in

complex scenarios. This is fundamental to avoid a high computation burden, allowing for a usable implementation in real-world analysis. Moreover, even if IPW provides an unbiased estimate under proper assumptions, we must also further investigate when IPW yields robust estimates to avoid a wrong definition of the nonparametric regression in the EVSI computation.

Once these steps are taken, we finally have to apply the definitive methodology to different realistic applications to demonstrate its potential in addressing real-world problems.

References

Ades, A., Lu, G., and Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling. *Medical decision making*, 24(2):207–227.

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, pages 211–217.

Assimon, M. M. (2021). Confounding in observational studies evaluating the safety and effectiveness of medical treatments. *Kidney360*, 2(7):1156–1159.

Benson, K. and Hartz, A. J. (2017). A comparison of observational studies and randomized, controlled trials. In *Research Ethics*, pages 213–221. Routledge.

Chesnaye, N. C., Stel, V. S., Tripepi, G., Dekker, F. W., Fu, E. L., Zoccali, C., and Jager, K. J. (2022). An introduction to inverse probability of treatment weighting in observational research. *Clinical kidney journal*, 15(1):14–20.

Claxton, K. P. and Sculpher, M. J. (2006). Using value of information analysis to prioritise health research: some lessons from recent uk experience. *Pharmacoeconomics*, 24:1055–1068.

Grimes, D. A. and Schulz, K. F. (2002). Bias and causal associations in observational research. *The lancet*, 359(9302):248–252.

Heath, A., Kunst, N., and Jackson, C. (2024). *Value of information for healthcare decision-making*. CRC Press.

Heath, A., Strong, M., Glynn, D., Kunst, N., Welton, N. J., and Goldhaber-Fiebert, J. D. (2022). Simulating study data to support expected value of sample information calculations: a tutorial. *Medical Decision Making*, 42(2):143–155.

Hegedus, E. J. and Moody, J. (2010). Clinimetrics corner: the many faces of selection bias. *Journal of Manual & Manipulative Therapy*, 18(2):69–73.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Rosenbaum, P. R., Rosenbaum, P., and Briskman (2010). *Design of observational studies*, volume 10. Springer.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Sackett, D. L. (1979). Bias in analytic research. In *The case-control study consensus and controversy*, pages 51–63. Elsevier.

Strong, M., Oakley, J. E., Brennan, A., and Breeze, P. (2015). Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Medical Decision Making*, 35(5):570–583.

Suissa, S. (2008). Immortal time bias in pharmacoepidemiology. *American journal of epidemiology*, 167(4):492–499.

Tripepi, G., Jager, K. J., Dekker, F. W., and Zoccali, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115(2):c94–c99.

Walker, A. M. (1996). Confounding by indication. *Epidemiology*, 7(4):335–336.

Chapter 5

Bayesian STEPP

5.1 Introduction

As we highlighted in the previous chapters, RCTs are the gold standard in health technology assessment (HTA), as they avoid different forms of bias through randomization. In most cases, when we want to compare the efficacy of two treatments, we consider the whole cohort of individuals to estimate the treatment effect.

However, in many cases, the treatment effect may be heterogeneous among individuals with different characteristics. In these cases, the 'one-size-fits-all' treatment recommendation should be discouraged, as we would lose important information necessary to ensure the cost-effectiveness of the treatment of interest (Willke et al. (2012)). Suppose, for instance, that a certain treatment has no relevant efficacy when we consider the general population, but this changes when we focus our attention on patients with a specific novel biomarker. Considering potential treatment-biomarker interactions has the potential for identifying subgroups of patients for whom the treatment is most effective or least effective.

As we have mentioned in §1.1, governments and other health-related payers have recognized that collecting additional evidence and properly assessing heterogeneity can be highly beneficial in restricting public subsidies to individuals for whom it is cost-effective in terms of net benefit (Coyle et al. (2003)). Subgroup analysis has recently been used to assess the value of heterogeneity, relying on the implementation of value of information for subgroup analyses (Espinoza et al. (2014)).

A very standard approach to analyze heterogeneity of treatment effect (HTE) is subgroup analysis (Rothwell (2005), Wang et al. (2007)). It involves partitioning the cohort of patients into disjoint subsets based on a specific covariate or risk factor and calculating the efficacy within each subgroup (Lagakos et al. (2006)).

As mentioned in §1.2.2.1, there are various principal possible drawbacks to this approach that we face when working with disjoint subsets: 1) randomization can't be ensured

if the sample size in subgroups is not large enough (Wang et al. (2010)), 2) lack of power due to smaller subgroups of patients (Lazar et al. (2016)), 3) categorizing patients by a single summary index can translate in losing important information on the effect of the baseline characteristic as a predictor of efficacy increases (Royston et al. (2006)), and 4) estimating differences in efficacy across different subgroups, one variable at a time, increases the probability of chance findings; this problem is known as *multiplicity* (Wang et al. (2007)). It can also happen that subgroups are defined according to some convenient, often arbitrary, cut-off points, which do not necessarily create clinically relevant subgroups and can lead to different types of bias (Lagakos et al. (2006), Wang et al. (2007)).

Predictive HTE analysis has been defined to address some of the above drawbacks (Kent et al. (2020)). It is performed by specifying a regression equation on RCT data with as predictors both the covariates that characterize the subgroups (risk factors), the exposure variable, and covariates-exposure interaction terms. This method allows for the simultaneous consideration of multiple relative characteristics in determining individualized differences in effectiveness. Despite this, predictive HTE still maintains potential problem of multiplicity, and it relies heavily on the correct definition of the outcome model. In this regard, recent machine-learning approaches provide a promising solution, as they can flexibly learn complex functional forms and interaction structures directly from the data, substantially reducing dependence on model specification (Hahn et al. (2020), Henderson et al. (2020),Glynn et al. (2024),Hu et al. (2021)). Nevertheless, these methods typically require large sample sizes to achieve stable and reliable estimates of individualized treatment effects.

Another method for examining possible treatment effect heterogeneity, the subpopulation treatment effect pattern plot (STEPP) method, was first introduced in (Bonetti and Gelber (2000)). STEPP is an exploratory graphical tool designed to help researchers investigate the potential heterogeneity of treatment effects and facilitate the interpreta-

tion of estimates of treatment effects derived from different but potentially overlapping subsets of patients. In addition, STEPP not only allows for graphical exploration but also has different significance-testing methods built in, enabling both visual exploration and robust statistical analysis (Bonetti and Gelber (2004)).

STEPP is primarily designed to evaluate HTE related to continuous covariates in a fashion that is similar to smoothing, by considering overlapping subgroups of patients; thus, borrowing strength from adjacent observations and obviating the need to define (often arbitrary) cutpoints to define subgroups. STEPP addresses some of the concerns associated with traditional subgroup analysis. Indeed, by considering overlapping subgroups, it mitigates the potential risks associated with a low sample size and its related drawbacks. Moreover, STEPP requires few assumptions, and treatment effects are computed using standard methods on well-defined groups of subjects. Finally, it provides a graphical display to show potentially complex interactions, thereby assisting researchers in interpreting the results (Bonetti et al. (2009)).

While traditional statistical methods for subgroup analysis divide the population into disjoint subgroups, STEPP takes a different approach by constructing overlapping subpopulations along the continuum of a continuous covariate of interest (e.g., a biomarker) or a baseline risk score, thereby improving the precision of the estimated treatment effect within the subgroups. There are several ways to divide the population into overlapping subgroups; the most common one is called the *sliding window method*. It works by specifying two parameters usually denoted as r_1 and r_2, with $r_1 < r_2 < n$. The former number, r_1, provides the approximate number of observations belonging to the overlap between a subpopulation and the next subpopulation as the sliding window moves from left to right. n represents the total number of observations. The latter parameter, r_2, indicates the (approximate) size of each subpopulation; after the specification of r_1 and r_2, the following algorithm generates the overlapping subgroups:

Sliding Window Algorithm for STEPP Subpopulation Construction

Let $\eta_1, \ldots, \eta_{\max}$ denote the ordered distinct values of the covariate $\tilde{X}$ that define the subgroups, with $\eta_{\max} = \sup\{\tilde{X}_i, i = 1, \ldots, n\}$ and $\tilde{X}_i$ the covariate value for patient i. Given fixed values r_1 and r_2 such that $r_1 < r_2 < n$, the subpopulations are constructed as follows:

1. Identify η_1^{upp} as the smallest η_t satisfying

$$\sum_{i=1}^{n} \mathbf{1}(\tilde{X}_i \leq \eta_t) \geq r_2.$$

   Set $\eta_1^{\text{low}} = \eta_0 = \eta_1 - 1$ and initialise $b = 2$.

2. For $b = 2, 3, \ldots$ slide across the range of $\tilde{X}$:

   (a) Find η_b^{low} as the smallest η such that

$$\sum_{i=1}^{n} \mathbf{1}(\tilde{X}_i \leq \eta_{b-1}^{\text{upp}}) \mathbf{1}(\tilde{X}_i > \eta_b^{\text{low}}) \leq r_1.$$

   (b) Find η_b^{upp} as the smallest η such that

$$\sum_{i=1}^{n} \mathbf{1}(\tilde{X}_i \leq \eta_b^{\text{upp}}) \mathbf{1}(\tilde{X}_i > \eta_b^{\text{low}}) \geq r_2.$$

   If no such value exists, set $\eta_b^{\text{upp}} = \eta_{\max}$ and stop.

As mentioned in Yip et al. (2016), the general guidelines for choosing r_1 and r_2 are:

1. choose r_2 large enough to obtain a good estimate of the treatment effect within subpopulations;
2. create at least $4 - 5$ subpopulations;
3. choose r_1/r_2 to be about $30\% - 50\%$ as the initial investigation;

4. make r_1, r_2 larger to obtain a smoother STEPP, but not so large to obtain fewer than four subpopulations;

5. assess the consistency of the result, a sensitivity analysis varying r_2 is recommended.

In addition, if covariate Z is continuous, if $(n - r2)/(r2 - r1)$ is close to an integer all the overlapping subgroups have approximately the same sample size. This avoids small last subgroup with associated low precision.

For each subpopulation, an estimate of treatment effect is computed. Such treatment effect estimates are clearly correlated, as neighboring subpopulations share some subjects. Therefore the asymptotic joint distribution is estimated and used to perform inference. Finally, the estimates are represented graphically in different diagrams to help the researcher interpret the results, together with the associated inference.

In this chapter, we introduce a Bayesian version of the STEPP method. As we will see in the next section, the Bayesian approach enables flexible modeling of the dependence between the relevant parameters related to the treatment effect and provides good control over the joint distribution of the parameters themselves and their uncertainty.

5.2 Bayesian STEPP (B-STEPP)

Suppose we face a random sample of n individuals, and that for each of them we observe an outcome Y, some covariates $\boldsymbol{X}$, and the exposure variable Z, with $z \in \{0, 1\}$, meaning that we are analyzing a two-treatment RCT. We assume that one of those covariates, $\tilde{X}$, is bounded and continuous; then, based on the value of $\tilde{X}$, we can divide the population into m overlapping subpopulations, which we denote by $S_1, \ldots, S_m$ (e.g., $\tilde{X}$ is a biomarker that is thought to modify the treatment effect). We let the outcome Y of individuals within the subpopulations depend on a vector parameters $\boldsymbol{\eta}^Z$, with $\boldsymbol{\eta}^{Z=1}$ representing the vector of parameters corresponding to the treatment group, and $\boldsymbol{\eta}^{Z=0}$ representing the vector of parameters corresponding to the control group. Also, let $\tilde{\boldsymbol{\eta}}^Z$ represent a

reparametrization of the vector of parameters $\boldsymbol{\eta}^Z$ used to express the form in which we compute treatment effects. We adopt this notation since treatment effects parameters are usually obtained as the difference of deterministic functions of the parameters $\boldsymbol{\eta}^Z$; i.e., $\tilde{\boldsymbol{\eta}}^Z = g(\boldsymbol{\eta}^Z)$, so that treatment effects can be computed as $\tilde{\boldsymbol{\eta}}^{Z=1} - \tilde{\boldsymbol{\eta}}^{Z=0}$. Note that this represents the difference between the two treatment-specific summaries of the two outcome distributions. For example, suppose we model survival time as log-normally distributed, that is, $t_Z|\eta, \sigma \sim lognorm(\eta^Z, \sigma^2)$, where the log-normal distribution is parameterized so that η^Z represents the mean of the distribution for group Z. If we aim to quantify the treatment effect in terms of the difference of average event rate (inverse of the mean), we must consider $\tilde{\eta}^Z = g(\eta^Z) = 1/\eta^Z$. We call the plot of the estimated subpopulation-specific outcome parameters of interest $\tilde{\boldsymbol{\eta}}^Z$, for $Z \in \{0, 1\}$ vs. the median values of $\tilde{X}$ within each subpopulation a "B-STEPP" plot, an example is the one of Figure §5.1. The same approach can be implemented for different definitions of treatment effect, such as the difference in survival probability at a fixed time point, or the hazard ratio under the assumption of proportional hazard, or other metrics.

As in the frequentist approach, the main question is how to deal with the dependence of the $\boldsymbol{\eta}^Z$ between subpopulations. As we will demonstrate in the next section, Bayesian modeling allows us to model this dependence in a highly flexible way. In particular, we will propose Bayesian approaches to the problem and apply them in different realistic scenarios.

We start by defining notation that will be useful throughout this work. We already mentioned that $\boldsymbol{\eta}^Z = (\eta^Z_{S_1}, \ldots, \eta^Z_{S_m})$ represent the vector of subpopulations outcome parameters, meaning that η_{S_i} is a relevant parameter of the outcome distribution within subpopulation S_i. Moreover, we assume that each subpopulation intersects only with the adjacent subpopulation(s), originating $2m-1$ disjoint subpopulations $D_1, \ldots, D_{2m-1}$ as represented in Figure 5.1. In this context, we define η^Z_i for $i = 1, \ldots, 2m-1$ as some relevant outcome parameters related to each of the $2m-1$ disjoint subpopulations

$D_1, \ldots, D_{2m-1}$, that are generated by the m overlapping subpopulations $S_1, \ldots, S_m$ we want to analyze, again as illustrated in Figure 5.1. In particular, $\eta_i^{Z=1}$ represents the parameter corresponding to the treatment group in subpopulation D_i, while $\eta_i^{Z=0}$ represents the parameter corresponding to the control group in subpopulation D_i. We assume for each η_i^Z to be governed by a proper prior distribution and for the parameters with odd index (the ones related to subpopulations m generated by the overlap between two subpopulations) $\{\eta_1^Z, \eta_3^Z, \ldots, \eta_{2k+1}^Z, \ldots, \eta_{2m-1}^Z\}$ to be independent.

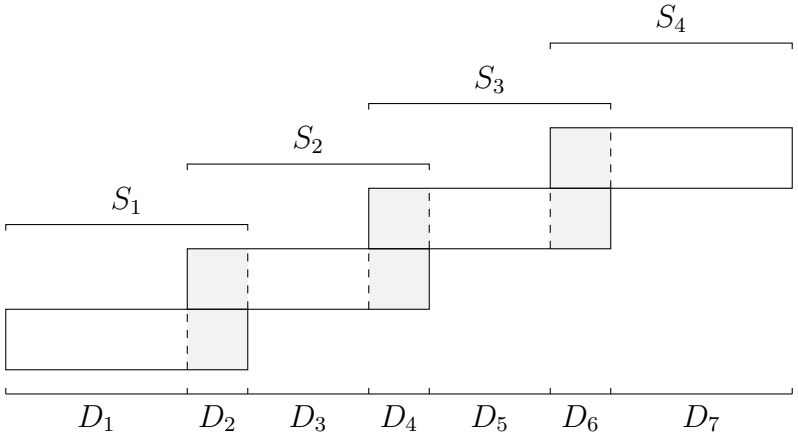


Figure 5.1: Illustration of four overlapping subpopulations, labeled $S_1, \ldots, S_4$ and the corresponding distinct subpopulations, labeled $D_1, \ldots, D_7$. The shaded regions indicate overlap. Here $m = 4$.

For the parameters with an even index (the ones related to subpopulations generated by the overlap between two subpopulations) $\{\eta_2^Z, \eta_4^Z, \ldots, \eta_{2k}^Z, \ldots, \eta_{2m-2}^Z\}$, since these parameters correspond to overlapping intervals, we assume they depend at least on the adjacent parameters. For example η_{2k}^Z depends on η_{2k-1}^Z and η_{2k+1}^Z. It becomes necessary to define a form for this dependence, which we do below in various ways. Note that Figure 5.1 assumes that the subsets have at most two intersections each. In general, each subset might have intersections with different subsets, e.g., S_1 intersects not only with S_2 but also with S_3. We will generalize the methods to more general patterns in Section 4.

One important remark is that throughout the work we will avoid specifying explicitly the indicator of group Z in $\boldsymbol{\eta}^Z$ if not strictly needed. Therefore, from now on, when

specifying explicitly Z is not necessary, we will just refer to $\boldsymbol{\eta}$, and $\tilde{\boldsymbol{\eta}}$, meaning that all the procedures are performed for the control and treatment groups separately.

5.2.1 Bayesian dependence modeling

The key ingredient to this implementation of STEPP is related to modeling the dependence between the different outcome parameters $\boldsymbol{\eta}$. In the frequentist setting, dependence is modeled explicitly by finding the joint asymptotic distribution of the estimators and performing inference based on it. In the Bayesian setting, we want to model the dependence of the relevant outcome parameters $\tilde{\boldsymbol{\eta}}$ by directly modeling the dependence between the parameters $\boldsymbol{\eta}$ that drive the model and affect the outcome. This is possible since $\boldsymbol{\eta}$ can be chosen so that the relevant outcome parameters $\tilde{\boldsymbol{\eta}}$, we want to compare to obtain treatment effects, are deterministic functions of the parameters $\boldsymbol{\eta}$ (or equal to $\boldsymbol{\eta}$).

First, we assume independence between the parameters related to the disjoint subsets $\{\eta_1, \eta_3, \ldots, \eta_{2k+1}, \ldots, \eta_{2m-1}\}$ and set a proper prior distribution for each of the parameters η_{2k+1} for $k = 1, .., m-1$; then, in this setting, the dependence between $\boldsymbol{\eta}$ is modeled by assuming the parameters related to the subpopulations generated by the overlap between two subsets $\{\eta_2, \eta_4, \ldots, \eta_{2k}, \ldots, \eta_{2m-2}\}$ to be linked with $\{\eta_1, \eta_3, \ldots, \eta_{2k+1}, \ldots, \eta_{2m-1}\}$ by different structural relationships:

$$\eta_{2k} = f(\boldsymbol{\gamma}_{2k}, \boldsymbol{\eta}_{[-2k]}) \quad k = 1, ..., m-1$$

where $\boldsymbol{\eta}_{[-2k]}$ is the set of parameters $\boldsymbol{\eta}$ except $\boldsymbol{\eta}_{2k}$ and $\boldsymbol{\gamma}$ are some parameters that could be random or constant. In the case of random $\boldsymbol{\gamma}$, we set a proper prior distribution.

We list some of the possible explicit definitions of f, that is, representations of dependence between the parameters:

- **Arithmetic mean**: we define η_{2k} as:

$$\eta_{2k} = \frac{\eta_{2k-1} + \eta_{2k+1}}{2}$$

- **Hierarchical convex combination**: we define η_{2k} as:

$$\eta_{2k} = \gamma_{2k}\eta_{2k-1} + (1 - \gamma_{2k})\eta_{2k+1}$$

  with γ modelled by a proper prior distribution, e.g. $\gamma_i \overset{iid}{\sim} U(0, 1)$.

- **General Hierarchical structure**: we define η_{2k} as:

$$\eta_{2k} = \gamma_{2k}\eta_{2k-1} + \beta_{2k}\eta_{2k+1}$$

  with γ and β modelled by two proper prior distributions, e.g. $\gamma_i \overset{iid}{\sim} U(0, 1)$ and $\beta_i \overset{iid}{\sim} U(0, 1)$.

Notice that the models above have a decreasing number of assumptions. The first assumes a deterministic relation between two consecutive subpopulations parameters (the arithmetic mean) to model dependence. In contrast, the second assumes a generalization of the arithmetic mean (convex combination) with convex coefficient defined as a random variable and estimated using MCMC. The last approach only assumes a linear combination, allowing for greater flexibility between consecutive subpopulations, with linear coefficients defined as random variables and estimated again using MCMC (Robert et al. (1999)).

Once the model and the prior distributions for $\{\eta_1, \eta_3, \ldots, \eta_{2k+1}, \ldots, \eta_{2m-1}\}$ have been chosen, we must specify the likelihood $p(Y|\boldsymbol{\eta})$ and how the parameters $\boldsymbol{\eta}$ are linked to the relevant outcome parameters $\tilde{\boldsymbol{\eta}}$, to fully describe the complete hierarchical Bayesian model:

$$Y_i|\boldsymbol{\eta} \sim P(\boldsymbol{\eta})$$

$$\eta_{2k-1} \overset{iid}{\sim} F(\gamma_{2k-1}) \quad \text{if } k \in \{1, 2, 3, .., m\}$$

$$\boldsymbol{\eta}_{2k} = f(\gamma_{2k}, \boldsymbol{\eta}_{[-2k]}) \quad k = 1, ..., m-1$$

$$\gamma_k \overset{iid}{\sim} \pi(a) \quad \text{if } k \in \{1, 2, 3, .., 2m-1\}$$

With P Likelihood distribution, F prior distributions for parameters η_{2k} with $k \in \{1, 2, 3, \ldots, m\}$ and π hyperprior. Note that if $\boldsymbol{\gamma}$ are not random, one does not need to set the prior distribution π and the model is not hierarchical. As we specified in §5.2, this model is defined analogously for both control and treatment groups.

5.2.2 Inference

Once we have specified these elements, we can sample from the posterior distribution of $\boldsymbol{\eta}$ and therefore from the posterior distribution of $\tilde{\boldsymbol{\eta}}$ since we have assumed $\tilde{\boldsymbol{\eta}}$ to be a deterministic function of $\boldsymbol{\eta}$, $\tilde{\boldsymbol{\eta}} = g(\boldsymbol{\eta})$.

Finally, we set the distribution of the final relevant outcome parameters we want to compare to determine treatment effect for subpopulations $\{S_1, \ldots, S_n\}$ as $\{\tilde{\eta}_1, \tilde{\eta}_3, \ldots, \tilde{\eta}_{2k+1}, \ldots, \tilde{\eta}_{2m-1}\}$:

$$\tilde{\eta}_{S_i} = \tilde{\eta}_{2i-1} \quad i = 1, .., m$$

As in the frequentist approach, also in the Bayesian version it is important to not only to characterize the variability around the estimates (which can be done by computing the relevant posterior distributions), but also to quantify the confidence we have in supporting that there is an actual difference between the overall treatment effect and the group-specific ones.

To this aim, in the original work on STEPP Bonetti and Gelber (2004), the following statistics to test the null hypotheses $\tilde{\eta}_{S_1}^{eff} = ...\tilde{\eta}_{S_m}^{eff}$ is introduced:

$$T = \sup_j \left\{ \frac{|\tilde{\eta}_{S_j}^{eff} - \tilde{\eta}_{ALL}^{eff}|}{\sigma_j}, \quad j = 1, ..., m \right\} \tag{5.1}$$

where $\tilde{\eta}_{S_i}^{eff} = \tilde{\eta}_{S_i}^{Z=1} - \tilde{\eta}_{S_i}^{Z=0}$ is the treatment effect on subpopulation S_i, $\tilde{\eta}_{ALL}^{eff} = \tilde{\eta}_{ALL}^{Z=1} - \tilde{\eta}_{ALL}^{Z=0}$ is the overall treatment effect on the entire sample, σ_j is the variance of $\tilde{\eta}_{S_j}^{eff} - \tilde{\eta}_{ALL}^{eff}$. In the original work, the asymptotic distribution of the T statistic above under the null hypothesis is computed, and the relevant p-value is calculated.

In our setting, to understand if there is an actual difference between group-specific and overall treatment, we will focus our attention on two relevant quantities. The first, T_j, is simply the difference between the group-specific treatments and the general one.

$$T_j = \tilde{\eta}_{S_j}^{eff} - \mu(\tilde{\eta}_{ALL}^{eff}) \quad j = 1, ..., m$$

Note that $\tilde{\eta}_{S_j}^{eff}$ is a random variable for which we will compute the posterior distribution. While $\mu(\tilde{\eta}_{ALL}^{eff})$ is the mean of the posterior distribution of $\tilde{\eta}_{ALL}^{eff}$, the random overall treatment effect. The posterior distribution of T_j provides the posterior probability that there is a difference between the group j treatment effect and the general one. For example, suppose that the posterior density of T_1 is the one in Figure 5.2:

Then, computing credible intervals enables us to calculate the probability that an actual difference exists. In particular, computing the tail probability $P(T_1 < 0)$, as shown in the figure above, provides the group-specific probability of no difference between group-specific and overall treatment effects.

A second quantity on which we can focus our attention is $\tilde{T}$, defined as:

$$\tilde{T} = \sup_j \{|\tilde{\eta}_{S_j}^{eff} - \mu(\tilde{\eta}_{ALL}^{eff})| \quad j = 1, ..., m\}$$

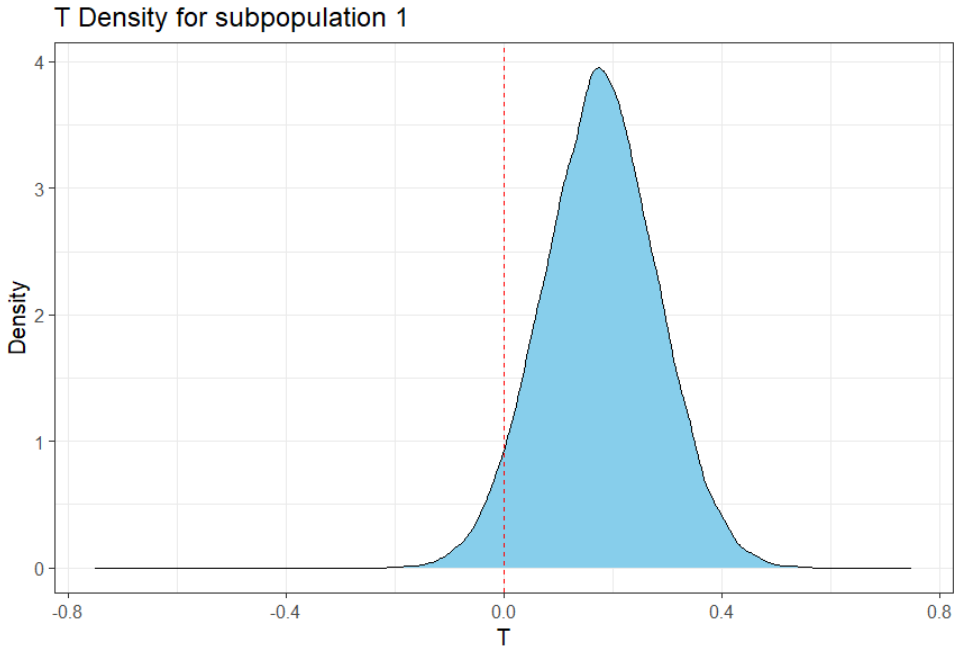


Figure 5.2: Posterior density of T_1 from application in §5.3.1, with red vertical dashed line with equation $T_1 = 0$.

Again, we can estimate (by MCMC) the posterior distribution of $\tilde{T}$ directly. Figure 5.3 represents the posterior density of $\tilde{T}$ coming from an application we will describe later.

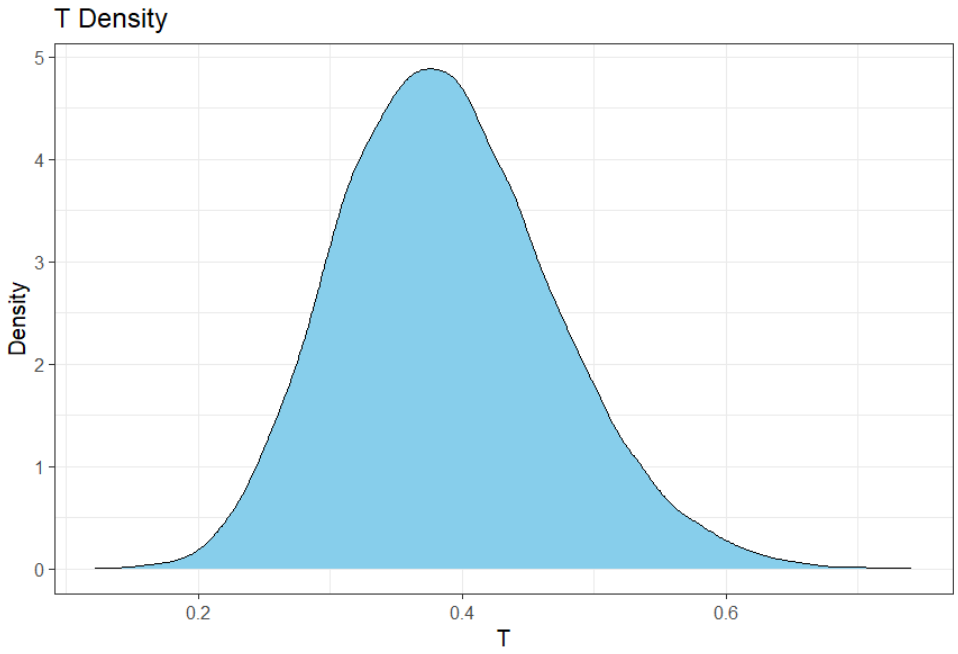


Figure 5.3: Posterior density of $\tilde{T}$ from application in §5.3.1.

This quantity provides information on the intensity of the difference between group-specific treatment effects and the overall treatment effect. In particular, $P(\tilde{T} < x)$ is the probability that all the group-specific treatment effects differ from the general one by less than x in absolute value. Note that this means that the probability that there exists at least one group whose specific treatment effect differs from the other groups-specific treatment for at least x is at least $1 - P(\tilde{T} < x)$. In application, one can proceed in two different ways; first, by finding x such that $P(\tilde{T} < x) = \alpha$ where α will be the credibility level we want to ensure in our analysis. Second, one may already want to test a specific value of x, maybe linked to some clinically meaningful level of the difference in effects, and compute the probability that all the group-specific treatment effects differ from the general one by less than x.

5.3 Applications

We apply our methodology to two different studies. In the first study, we focus on a binary outcome/endpoint, while the second one consists of a study in a survival setting with a continuous endpoint.

5.3.1 The Aspirin/Folate Polyp Prevention Study

The Aspirin/Folate Polyp Prevention Study was a randomized, double-blind, placebo-controlled trial of the efficacy of oral aspirin, folic acid, or both to prevent colorectal adenomas in individuals having a prior history of such lesions (Baron et al. (2003)). The initial findings from the aspirin analysis indicated that low-dose aspirin was associated with a moderate chemopreventive effect on colorectal adenomas.

The analysis of this study with a frequentist STEPP approach was proposed in (Yip et al. (2016)), where the authors focused their analyses on the aspirin component of the study with the presence of adenomas as the endpoint. There were 1121 participants

randomized to three aspirin groups (placebo, 81 mg/day, and 325 mg/day). Participants were followed for 3 years and then underwent colonoscopy. The primary endpoint was the occurrence of any pathologically confirmed adenomas. A total of 1084 participants underwent colonoscopy follow-up at 3 years. In (Yip et al. (2016)), the authors employ STEPP to investigate the presence of HTE in placebo vs 81 mg/day groups, in patients of different ages, and relevant differences in individuals between 53 and 61 years old.

We first divide our population into $m = 10$ overlapping intervals with respect to the age of the patients using the standard sliding window approach, setting $r_1 = 30$ and $r_2 = 100$, thus obtaining 19 disjoint subsets $D_1, \ldots, D_{19}$. Then, we model the different variables of the study as follows:

$$Y_i|\eta_j \overset{iid}{\sim} \text{Bern}(\eta_j), \quad j \in \{1, 2, \ldots, 19\}$$

$$\eta_{2k-1} \overset{iid}{\sim} \text{Beta}(1, 1), \quad k \in \{1, 2, 3, .., 10\}$$

$$\eta_{2k} = \gamma_{2k}\eta_{2k-1} + (1 - \gamma_{2k})\eta_{2k+1}, \quad k \in \{1, 2, 3, .., 9\}$$

$$\gamma_{2k} \overset{iid}{\sim} U(0, 1), \quad k \in \{1, 2, 3, .., 9\}$$

Note that we are considering the *hierarchical convex combination* from the previous section. Moreover, one can also note that, in this case, the risk of exhibiting an adenoma is modeled by η itself, hence $\tilde{\eta} = g(\eta) = \eta$. Again, as we specified in §5.2, the model above is defined analogously for both control and treatment groups.

We sample from the posterior distribution of η using MCMC (Robert et al. (1999)), and plot both $\tilde{\eta}_{S_i}^{Z=1}$ and $\tilde{\eta}_{S_i}^{Z=0}$ for the 10 subpopulations that overlap, obtaining the results shown in Figure 5.4.

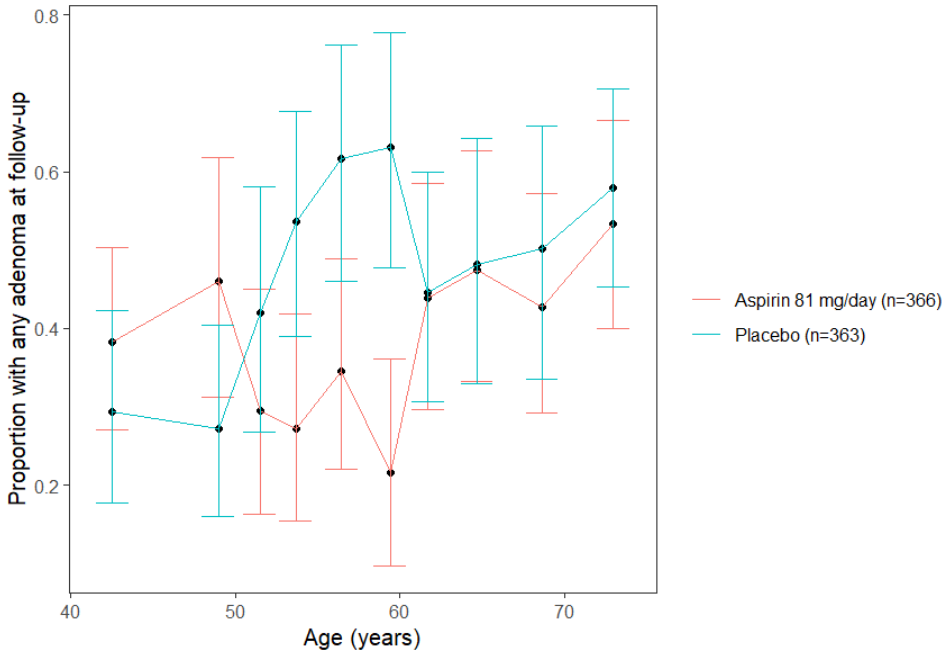


Figure 5.4: B-STEPP plot for the Aspirin/Folate Polyp Prevention Data. 81 mg/day (blue) vs. placebo groups (red), with 90% credible intervals of the marginal posterior distributions.

The vertical intervals represent the 90% credible intervals of the marginal posterior distributions. As we mentioned above, in applications, sensitivity analysis on the value of r_1 and r_2 is recommended. Therefore, we add a preliminary sensitivity analysis in Figure 5.5 which preliminary confirms the stability of our results.

We represent in Figure 5.6 the posterior densities of T_j, $j = 1, \ldots, 10$ to test the hypothesis that there is no difference in treatment effect between each subpopulation and the overall study population.

	$P(T_1 < 0)$	$P(T_2 < 0)$	$P(T_3 > 0)$	$P(T_4 > 0)$	$P(T_5 > 0)$	$P(T_6 > 0)$	$P(T_7 < 0)$	$P(T_8 < 0)$	$P(T_9 < 0)$	$P(T_{10} < 0)$
Tail probability	0.05	0.01	0.40	0.08	0.08	0.01	0.27	0.26	0.46	0.37

Table 5.1: Tail probabilities of the posterior distributions of T_j, $j = 1, \ldots, 10$.

From the posterior densities in Figure 5.6, it appears that groups 1, 2, and 6 are more likely to have a real difference between their group-specific treatment effect and the general one. Additionally, groups 5 and 6 seem to exhibit the same real difference, just

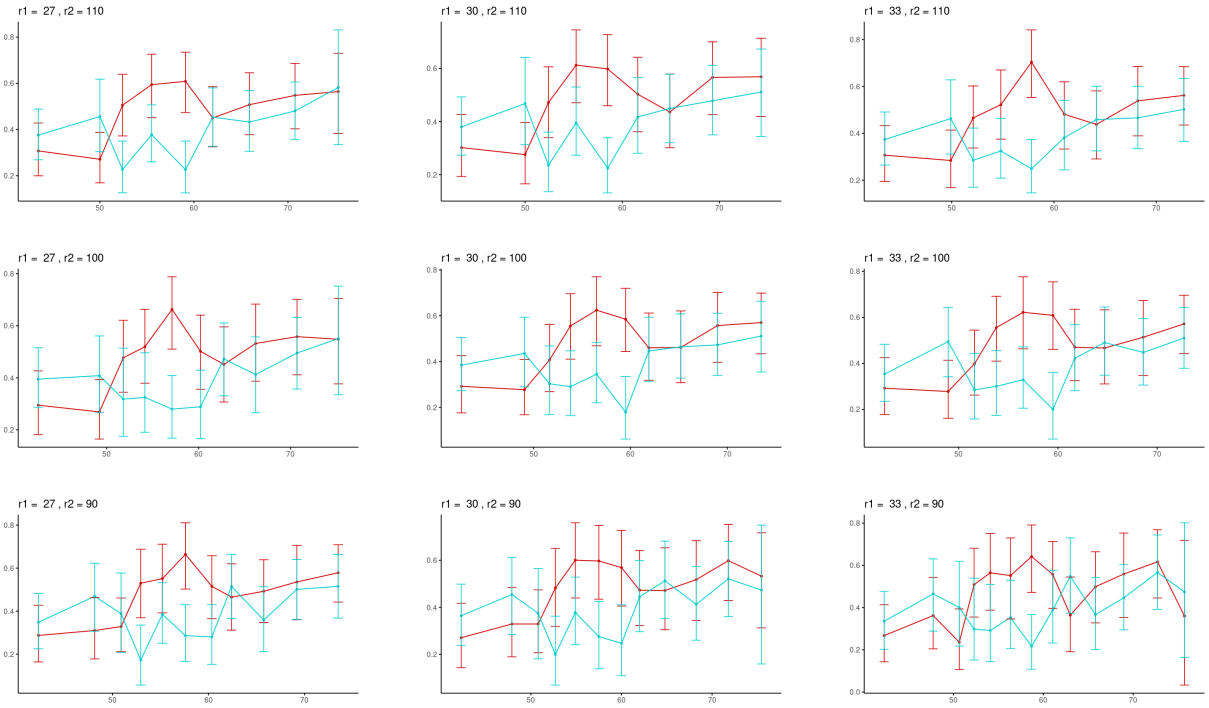


Figure 5.5: Sensitivity analysis for $r_1 = \{27, 30, 33\}, r_2 = \{90, 100, 110\}$

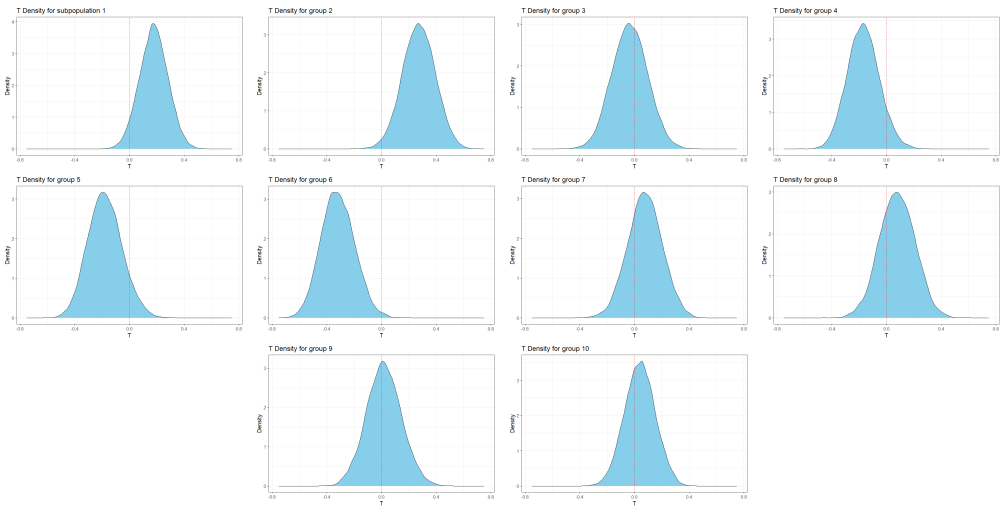


Figure 5.6: Posterior densities of $T_j = \tilde{\eta}_{S_j} - \mu(\tilde{\eta}_{ALL})$, with red vertical dashed line with equations $T_j = 0$, for $j = 1, \ldots, 10$ in Aspirin/Folate Polyp Prevention Study.

with a slightly lower probability. From these results, we can say that Bayesian STEPP analysis largely confirms the prior analysis, showing that the aspirin benefit appears to be largest in the participants who are in their mid-to-late 50's in age. Note that, in Table 5.1, we also computed the tail probabilities of the posterior distributions of T_j, $j = 1, \ldots, 10$,

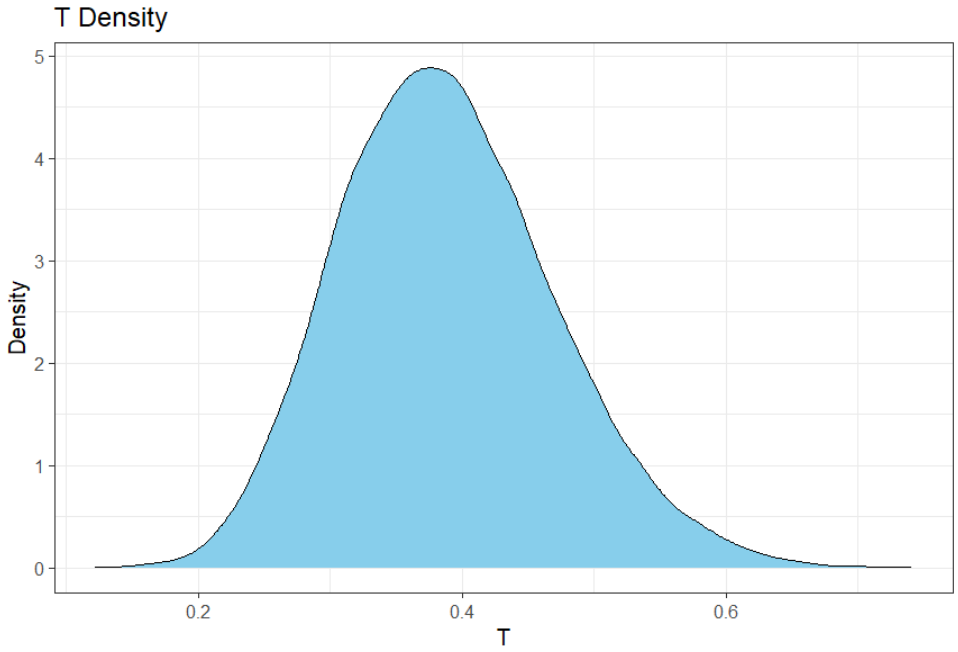


Figure 5.7: Posterior distribution of the $\tilde{T}$ statistic

and therefore, the related credible intervals, which confirms the conclusions drawn from the T_j statistics posterior distributions.

We may also plot the posterior distribution of the statistic $\tilde{T}$ (see Figure 5.7). As we mentioned above, by computing $P(\tilde{T} < x)$, we can calculate the probability that all the groups-specific treatment effects differ from the general one by less than x in absolute value. In this case, we can compute $P(\tilde{T} < 0.27) = 0.05$; therefore, there is only 5% probability that all the group-specific treatment effects differ from the general one by less than 0.27 in absolute value.

5.3.2 The NSABP B-20 Breast Cancer Clinical Trial

The B-20 study of the National Surgical Adjuvant Breast and Bowel Project (NSABP) (Fisher et al. (1997)) was conducted to determine whether chemotherapy plus tamoxifen would be of greater benefit than tamoxifen alone in the treatment of patients with breast cancer positive for estrogen receptors in the axillary lymph nodes.

The dataset is composed of 2363 individuals with recorded DFS (disease-free survival time), the censoring indicator, the exposure variable (whether they receive tamoxifen only or tamoxifen + chemotherapy), the ER value and the age of the individual.

We first divide our population into $m = 8$ overlapping intervals with respect to the age of the patients using the standard sliding window approach, with $r_1 = 60$ and $r_2 = 300$, obtaining 15 disjoint subsets $D_1, \ldots, D_{15}$. We implement both an exponential model and a lognormal survival model

5.3.3 Exponential model

First, we consider a Bayesian exponential model as follows:

$$t_i|\eta_j \sim \exp(\eta_j), \quad j \in \{1, 2, \ldots, 15\}$$

$$\eta_{2k-1} \overset{iid}{\sim} \text{gamma}(1, 1) \quad k \in \{1, 2, 3, .., 8\}$$

$$\eta_{2k} = \gamma_{2k}\eta_{2k-1} + \beta_{2k}\eta_{2k+1} \quad k \in \{1, 2, 3, .., 7\}$$

$$\gamma_{2k} \overset{iid}{\sim} U(0, 1) \quad \text{if } k \in \{1, 2, 3, .., 7\}$$

$$\beta_{2k} \overset{iid}{\sim} U(0, 1) \quad \text{if } k \in \{1, 2, 3, .., 7\}$$

In this case, we are considering the *general hierarchical structure* from the previous section. Moreover, one can also note that, in this case, the average event rate of DFS (which is also the level of the constant hazard function in the case of the exponential survival model) is modeled by η itself, hence $\tilde{\eta} = \eta$. Again, as we specified in §5.2, the model above is defined analogously for both control and treatment groups.

We sample from the posterior of η using MCMC, and plot both $\tilde{\eta}_{S_i}^{Z=1}$ and $\tilde{\eta}_{S_i}^{Z=0}$, which represent the relevant outcome parameters for the $m = 8$ overlapping subpopulations, obtaining the result shown in Figure 5.8. As in the previous application, the vertical

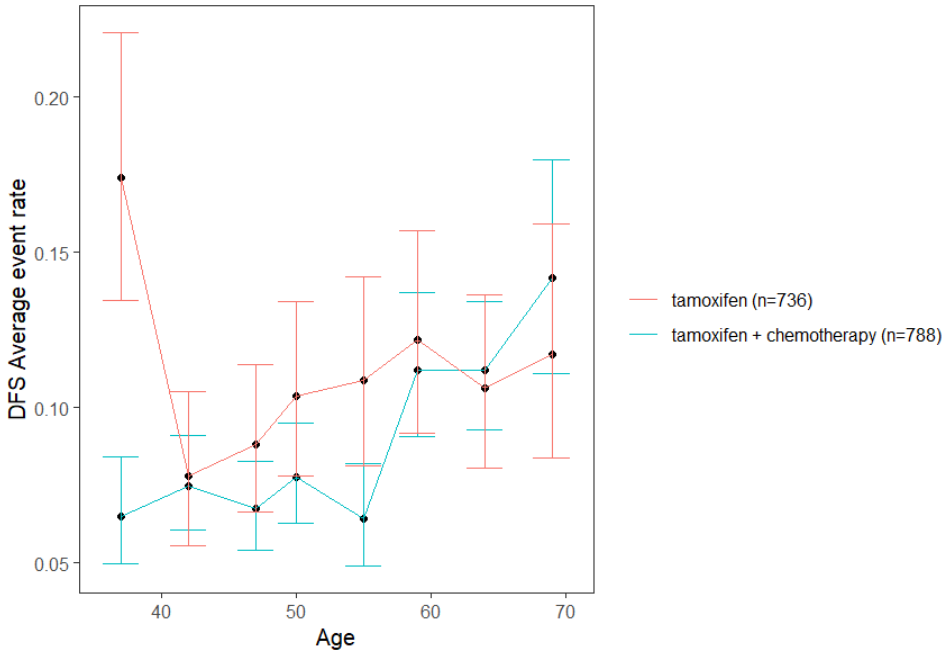


Figure 5.8: B-STEPP plot for the NSABP B-20 study (exponential survival model). Tamoxifen (red) vs tamoxifen + chemotherapy (blue), with 90% credible intervals of the marginal posterior distributions.

intervals represent the 90% credible intervals.

We then proceed as above, plotting the posterior densities of the T_j for each subpopulation (see Figure 5.9).

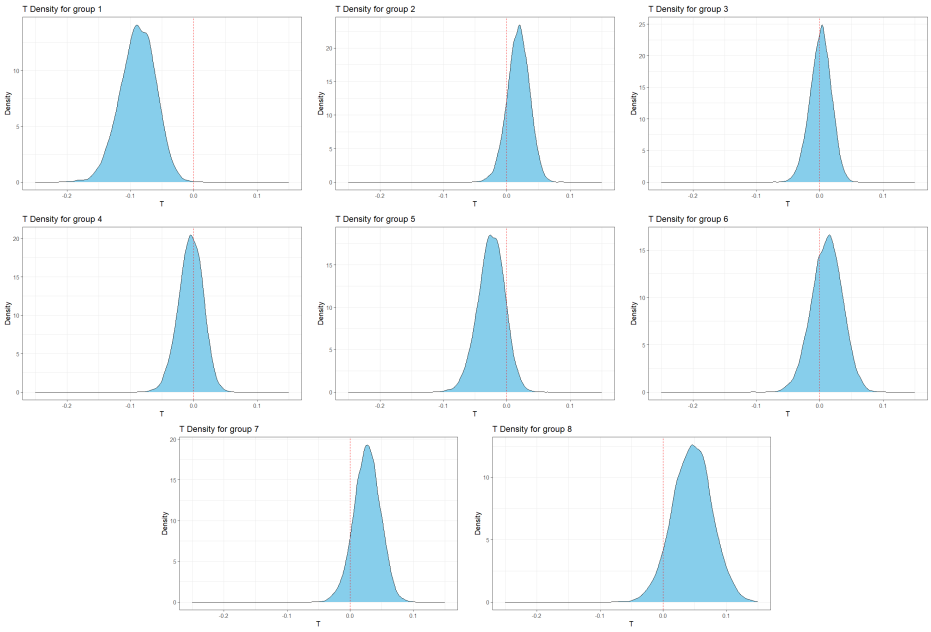


Figure 5.9: Posterior densities of $T_j = \tilde{\eta}_{S_j} - \mu(\tilde{\eta}_{ALL})$, with red vertical dashed line with equations $T_j = 0$, for $j = 1, \ldots, 8$ in NSABP B-20 study with exponential model.

	$P(T_1 > 0)$	$P(T_2 < 0)$	$P(T_3 < 0)$	$P(T_4 > 0)$	$P(T_5 > 0)$	$P(T_6 < 0)$	$P(T_7 < 0)$	$P(T_8 < 0)$
Tail probability	0.00	0.14	0.47	0.40	0.13	0.32	0.10	0.08

Table 5.2: Tail probabilities of the posterior distributions of T_j, $j = 1, \ldots, 8$.

From the densities in Figure 5.9, it seems that the group for which it is more likely that there is a real difference between group-specific treatment effects and the general one is group 1 (median age of 37 years). Again, we report the tail probabilities for each of the posterior distributions of T_j, $j = 1, \ldots, 10$ to compute the credible intervals and quantify the relevant probabilities (see Table 5.2).

In Figure 5.10, we show the posterior distribution of the statistic $\tilde{T}$.

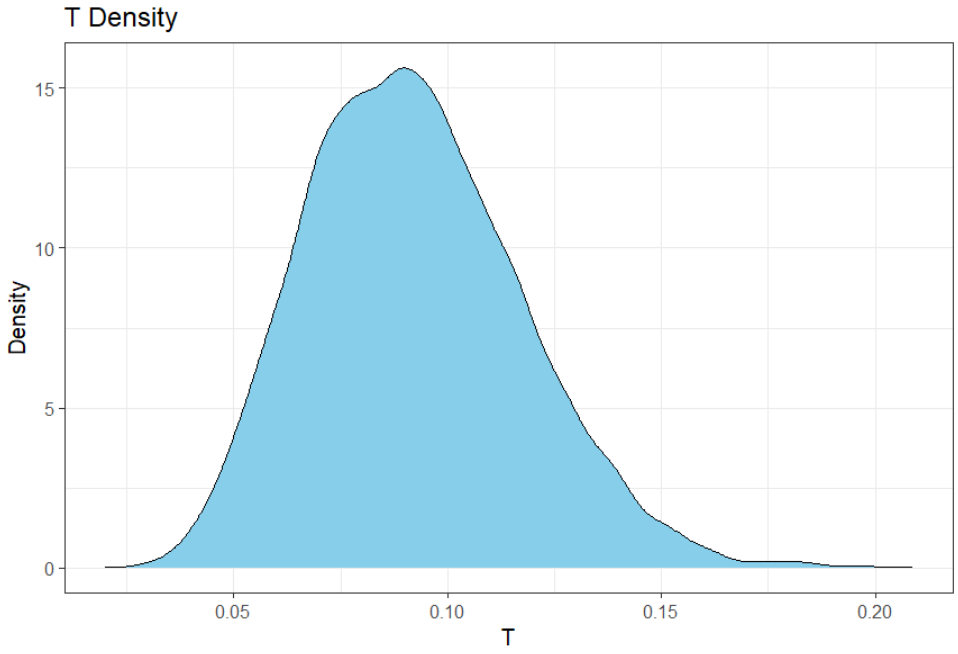


Figure 5.10: Posterior distribution of $\tilde{T}$, for the NSABP B-20 study (exponential survival model)

In this case, $P(\tilde{T} < 0.055) = 0.05$; therefore, there is a 5% probability that all the group-specific treatment effects differ from the general one by less than 0.055 in absolute value.

5.3.4 Lognormal model

In the same context as above, we consider a Lognormal distribution to model the survival data.

$$t_i|\mu_j, \sigma_j^2 \sim \text{lognorm}(\eta_j, \sigma_j^2), \quad j \in \{1, 2, \ldots, 15\}$$

$$\eta_{2k-1} \overset{iid}{\sim} \text{gamma}(8, 1) \quad \text{if } k \in \{1, 2, 3, .., 8\}$$

$$\sigma_k^2 \overset{iid}{\sim} Gamma(3, 1) \quad k = 1, .., 15$$

$$\eta_{2k} = \gamma_{2k}\eta_{2k-1} + \beta_{2k}\eta_{2k+1} \quad \text{if } k \in \{1, 2, 3, .., 7\}$$

$$\gamma_{2k} \overset{iid}{\sim} U(0, 1) \quad \text{if } k \in \{1, 2, 3, .., 7\}$$

$$\beta_{2k} \overset{iid}{\sim} U(0,1) \quad \text{if } k \in \{1,2,3,..,7\}$$

In this context, we reparametrize the lognormal distribution with respect to the mean η and variance σ^2. Again, as we specified in §5.2, the model above is defined analogously for both control and treatment groups.

In this case, to allow a better comparison with the results of the exponential model, we want to show again the results related to the average event rate of DSF across subpopulations; to this aim, we set $\tilde{\boldsymbol{\eta}} = 1/\boldsymbol{\eta}$. Figure 5.11 shows the corresponding B-STEPP plot. As in the previous applications, the vertical intervals represent the 90% credible

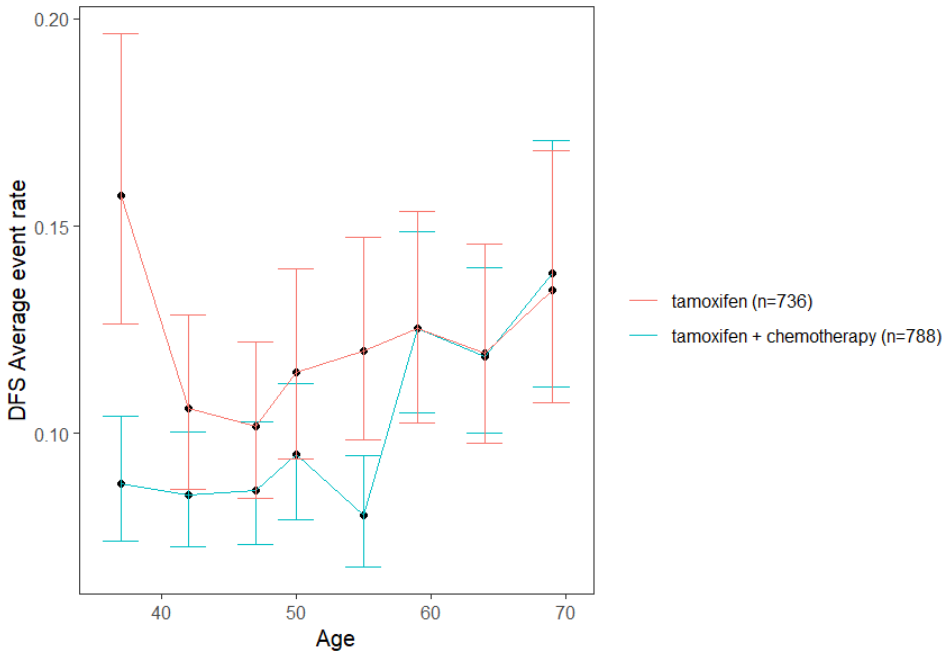


Figure 5.11: B-STEPP plot for the NSABP B-20 study (lognormal survival model). Tamoxifen (red) vs tamoxifen + chemotherapy (blue), with 90% credible intervals of the marginal posterior distributions.

intervals.

Again, we can plot the posterior densities of the T_j for each subpopulation (see Figure 5.12) and compute the relevant tail probabilities (see Table 5.3).

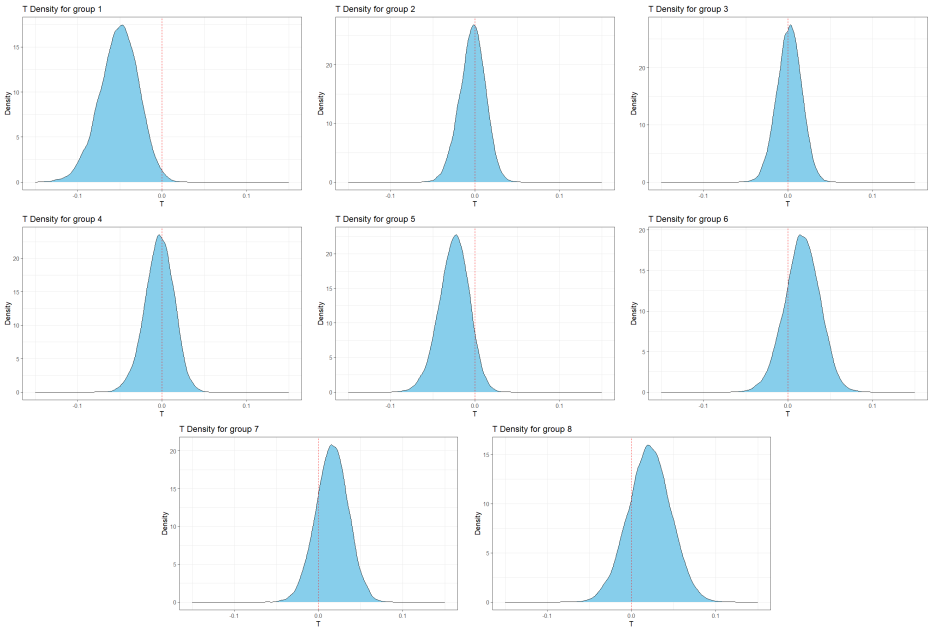


Figure 5.12: Posterior densities of $T_j = \tilde{\eta}_{S_j} - \mu(\tilde{\eta}_{ALL})$, with red vertical dashed line with equations $T_j = 0$, for $j = 1, \ldots, 8$ in NSABP B-20 study with lognormal model.

	$P(T_1 > 0)$	$P(T_2 < 0)$	$P(T_3 < 0)$	$P(T_4 > 0)$	$P(T_5 > 0)$	$P(T_6 < 0)$	$P(T_7 < 0)$	$P(T_8 < 0)$
Tail probability	0.01	0.41	0.46	0.43	0.07	0.22	0.17	0.21

Table 5.3: Tail probabilities of the posterior distributions of T_j, $j = 1, \ldots, 8$.

Just as in the exponential model above, group 1 is again the most likely group to exhibit a relevant difference between group-specific treatment effects and the general one. The results here also suggest a possible relevant difference in group 5. Note how these results are in agreement with the results shown in Fisher et al. (1997), while allowing for a finer granularity in the heterogeneity induced by age. In particular, the subpopulations that benefit the most from the treatment are the ones with age below 40 (very young people) and the subpopulation around 55 (middle-aged people).

Finally, we may plot the posterior distribution of the statistic $\tilde{T}$ (see Figure 5.13). In this case, we can only ensure that there is a 5% probability that all the group-specific treatment effects differ from the general one by less than 0.024 in absolute value, since $P(\tilde{T} < 0.024) = 0.05$.

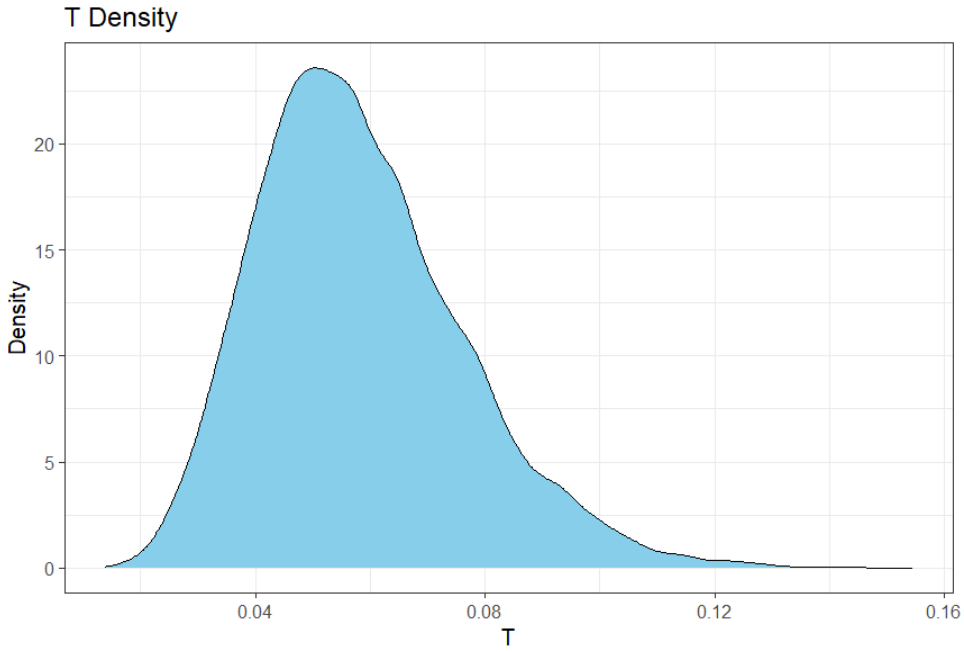


Figure 5.13: Posterior distribution of $\tilde{T}$, for the NSABP B-20 study (lognormal survival model)

5.4 Simulation study for model robustness

In this section, we conduct a preliminary exploration of B-STEPP's performance in estimating subpopulation parameters η_{S_i}, $i = 1, \ldots, m$ by comparing it with crude frequentist estimates of the relevant parameters under different assumptions including model misspecification. We test the performance of B-STEPP in a survival context, where we want to estimate the hazard rate within subpopulations. In this case, the frequentist estimate of the hazard rate, ignoring the intersections, is computed as:

$$\hat{\eta}_{S_i} = \frac{\text{Number of Events in } D_{2i-1}}{\text{Total Person-Time in } D_{2i-1}} \quad i = 1, \ldots, m$$

Which is the subpopulation-specific maximum Likelihood estimator of η_{S_i} when one assumes exponential survival.

For the B-STEPP method, we adopt the Gamma-Exponential model introduced above

and use the posterior mean of the exponential parameter η as the estimator of the hazard rate, since η directly represents the hazard rate in the exponential distribution.

We generate data assuming different possible ways of modeling dependencies between subpopulations' parameters and under different data-generating mechanisms. Specifically, the following tables compare B-STEPP with crude estimates in terms of absolute errors between the correct parameters and the estimated ones. We compare these quantities under different percentages of overlapping, i.e. percentages of individuals belonging to the intersections of subsets.

In the first scenario, we assume data to be generated by an exponential model $\eta_{2k-1} \overset{iid}{\sim} Gamma(1,1),\ k \in \{1,2,3,..,5\}$ and dependence modeled by the *arithmetic mean* relation introduced above; here, we use the same model to make inferences. We consider the same structure of Figure 5.1 with $m = 5$ different sets $S_1, .., S_m$. We assume a population of 150 individuals and run simulations 100 times for different percentages of overlap, i.e., the proportion of S_i and S_{i+1} for $i = 1, \ldots, m$ that overlap. We compute the mean squared error (MSE) and its decomposition in variance and squared bias; table 5.4 shows the results.

	prop. overlap 0.15	prop. overlap 0.25	prop. overlap 0.35
STEPP MSE	0.0865 (0.0859 ; 0.0006)	0.1205 (0.1201 ; 0.0005)	0.1085 (0.1080 ; 0.0006)
crude estimate MSE	0.1254 (0.1228 ; 0.0026)	0.1914 (0.1872 ; 0.0042)	0.2549 (0.2460 ; 0.0089)

Table 5.4: Comparison between *arithmetic mean* B-STEPP estimates and crude estimates in terms of MSE with varying proportion of overlap, variance and squared bias in brackets. Dependence is modeled as the arithmetic mean between consecutive subsets, and data generated from
$\eta_{2k-1} \overset{iid}{\sim} Gamma(1,1),\ k \in \{1,2,3,..,5\}$.

As we can see the B-STEPP produces better estimates than the crude ones. This is

mainly driven by the high variance of the crude estimates which B-STEPP addresses by allowing the different subsets to overlap.

After that, we define the data-generating mechanism so that the parameters definition in the intersections is not the arithmetic mean:

$$\eta_{2k} = \frac{\eta_{2k-1} + \eta_{2k+1}}{2} \quad k = 1, 2, 3, 4$$

but rather the product of the parameters related to two consecutive sets:

$$\eta_{2k} = \eta_{2k-1}\eta_{2k+1} \quad k = 1, 2, 3, 4 \tag{5.2}$$

Again, we assume data to be generated by an exponential model $\eta_{2k-1} \overset{iid}{\sim} Gamma(1, 1)$ if $k \in \{1, 2, 3, .., 5\}$. In this case, we apply both the *Arithmetic mean* model and the *Hierarchical* model to test the robustness of the methods to model misspecification. Table 5.5 shows the results in terms of MSE with variance and squared bias in brackets:

	prop. overlap 0.15	prop. overlap 0.25	prop. overlap 0.35
STEPP MSE	0.1322 (0.1286 ; 0.0036)	0.1270 (0.1143 ; 0.0127)	0.3424 (0.3130 ; 0.0294)
STEPP hier. MSE	0.1231 (0.1124 ; 0.0107)	0.1176 (0.1058 ; 0.0118)	0.1808 (0.1617 ; 0.0191)
crude estimate MSE	0.1686 (0.1622 ; 0.0064)	0.1670 (0.1623 ; 0.0048)	0.5782 (0.5682 ; 0.0101)

Table 5.5: Comparison between *arithmetic mean* B-STEPP estimates, *general hierarchical* B-STEPP estimates, and crude estimates in terms of MSE, variance and squared bias in brackets, with varying proportion of overlap. Dependence is modeled as in (5.2), and data generated from
$\eta_{2k-1} \overset{iid}{\sim} Gamma(1, 1),\ k \in \{1, 2, 3, .., 5\}$.

As we can see, the hierarchical model is much more robust than the classical arithmetic mean model, and it is also more robust than the crude estimates that do not account for

parameters' dependence; and, as above, have a high variance.

Finally, maintaining data generated by an exponential model $\eta_{2k-1} \overset{iid}{\sim} Gamma(1,1)$ if $k \in \{1,2,3,..,5\}$, we define the parameters in the intersection in the data-generating mechanism as:

$$\eta_{2k} = \left| 1 - \frac{\eta_{2k-1} + \eta_{2k+1}}{2} \right| \quad k = 1,2,3,4 \tag{5.3}$$

We obtain the results reported in Table 5.6.

	prop. overlap 0.15	prop. overlap 0.25	prop. overlap 0.35
STEPP MSE	0.1372 (0.1164 ; 0.0208)	0.2250 (0.1722 ; 0.0528)	0.4246 (0.3208 ; 0.1039)
STEPP hier. MSE	0.1063 (0.1007 ; 0.0056)	0.1042 (0.0990 ; 0.0053)	0.1904 (0.1879 ; 0.0025)
crude estimate MSE	0.1686 (0.1622 ; 0.0064)	0.1670 (0.1623 ; 0.0048)	0.5782 (0.5682 ; 0.0101)

Table 5.6: Comparison between *arithmetic mean* B-STEPP estimates, *general hierarchical* B-STEPP estimates, and crude estimates in terms of MSE, variance and squared bias in brackets, with varying proportion of overlap. Dependence is modeled as in (5.3), and data generated from
$\eta_{2k-1} \overset{iid}{\sim} Gamma(1,1),\ k \in \{1,2,3,..,5\}$.

In this context, we observe a behavior similar to that of the previous model.

Lastly, we test the robustness of the methods in terms of model misspecification related to the data-generating mechanism of the parameters in the disjoint subsets. In particular, we do not simulate data assuming that η follow the same distribution we defined in the Bayesian model above, i.e. $\eta_{2k-1} \overset{iid}{\sim} Gamma(1,1)$ for $k \in \{1,2,3,..,5\}$, but rather assume

$$\eta_k \sim \text{gamma}(1,1) \quad k = 1,5,9 \tag{5.4}$$

$$\eta_k \sim \text{gamma}(5,1) \quad k = 3,7$$

Moreover, we assume the dependencies between parameters to exhibit the following form.

$$\eta_{2k} = 0.9\eta_{2k-1} + 0.1\eta_{2k+1} \quad k = 1,2,3,4 \tag{5.5}$$

We compare the hierarchical method with the crude estimate, and obtain the following results:

	prop. overlap 0.15	prop. overlap 0.25	prop. overlap 0.35
STEPP hier. MSE	0.6466 (0.6248 ; 0.0218)	0.8180 (0.8029 ; 0.0151)	0.8251 (0.8133 ; 0.0118)
crude estimate MSE	1.0232 (1.0144 ; 0.0088)	1.6409 (1.6132 ; 0.0277)	2.4472 (2.3722 ; 0.0751)

Table 5.7: Comparison between *arithmetic mean* B-STEPP estimates, *general hierarchical* B-STEPP estimates, and crude estimates in terms of MSE, variance and squared bias in brackets, with varying proportion of overlap. Dependence is modeled as in (5.5), and data generated from (5.4).

We can observe that, also in face of data-generating model misspecification, B-STEPP leads to much more precise and robust estimates than crude estimates.

5.5 General patterns of subpopulations

Looking back at Figure 5.1, we note that we allowed only two consecutive intervals to overlap. In general, we may have n consecutive intervals with an intersection; in this case, we have to explicitly model the dependence structure for the individuals belonging to each intersection.

Suppose, for instance, that the STEPP structure is the one in Figure 5.14.

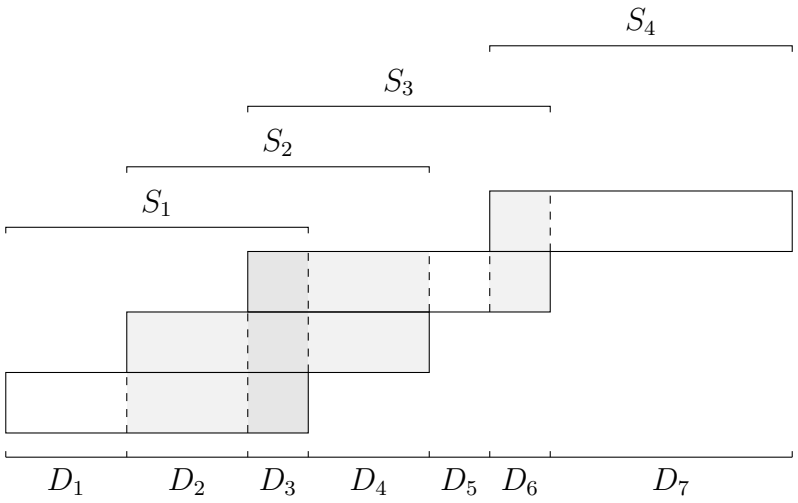


Figure 5.14: Illustration of four overlapping subpopulations, labeled $S_1, \ldots, S_4$ and the corresponding distinct subpopulations, labeled $D_1, \ldots, D_7$. The shaded regions indicate overlap

In this case, the models above can be modified as follows:

- **Arithmetic mean**: we define $\boldsymbol{\eta}_{2k}$, with $k = 1, \ldots, m-1$ as:

$$\boldsymbol{\eta}_k = \frac{\boldsymbol{\eta}_{S_{k-1}} + \boldsymbol{\eta}_{S_k} + \boldsymbol{\eta}_{S_{k+1}}}{3} \quad \text{for } D_k \text{ intersection of 3 subsets}$$

$$\boldsymbol{\eta}_k = \frac{\boldsymbol{\eta}_{S_{k-1}} + \boldsymbol{\eta}_{S_{k+1}}}{2} \quad \text{for } D_k \text{ intersection of 2 subsets}$$

- **Hierarchical convex combination**: we define $\boldsymbol{\eta}_k$ as:

$$\boldsymbol{\eta}_k = \alpha_k \boldsymbol{\eta}_{S_{k-1}} + \beta_k \boldsymbol{\eta}_{S_k} + (1 - \alpha_k - \beta_k)\boldsymbol{\eta}_{S_{k+1}} \quad \text{for } D_k \text{ intersection of 3 subsets}$$

$$\boldsymbol{\eta}_k = \alpha_k \boldsymbol{\eta}_{S_{k-1}} + (1 - \alpha_k)\boldsymbol{\eta}_{S_{k+1}} \quad \text{for } D_k \text{ intersection of 2 subsets}$$

with γ modelled by a proper prior distribution, e.g. $\alpha_i, \beta_i \overset{iid}{\sim} U(0,1)$.

- **General Hierarchical structure**: we define $\boldsymbol{\eta}_{2k}$, with $k = 1, \ldots, m-1$ as:

$$\boldsymbol{\eta}_k = \alpha_k \boldsymbol{\eta}_{S_{k-1}} + \beta_k \boldsymbol{\eta}_{S_k} + \gamma_k \boldsymbol{\eta}_{S_{k+1}} \quad \text{for } D_k \text{ intersection of 3 subsets}$$

$$\boldsymbol{\eta}_k = \alpha_k \boldsymbol{\eta}_{S_{k-1}} + \beta_k \boldsymbol{\eta}_{S_{k+1}} \quad \text{for } D_k \text{ intersection of 2 subsets}$$

with α, γ and β modelled by two proper prior distributions, e.g. $\alpha_i \overset{iid}{\sim} U(0,1)$, $\gamma_i \overset{iid}{\sim} U(0,1)$ and $\beta_i \overset{iid}{\sim} U(0,1)$.

Again, as we specified in §5.2, the model above is defined analogously for both control and treatment groups.

To illustrate the results in the generalized scenario above we focus on the Aspirin/Folate Polyp study, and consider the scenario in which we have obtained $m = 23$ overlapping intervals $S_1, ..., S_{23}$ and obtaining 66 disjoint subsets: $D_1, ..., D_{66}$. Figure 5.14 represents the B-STEPP plot obtained by applying the *Arithmetic mean* assumption to model dependence with a general pattern of populations.

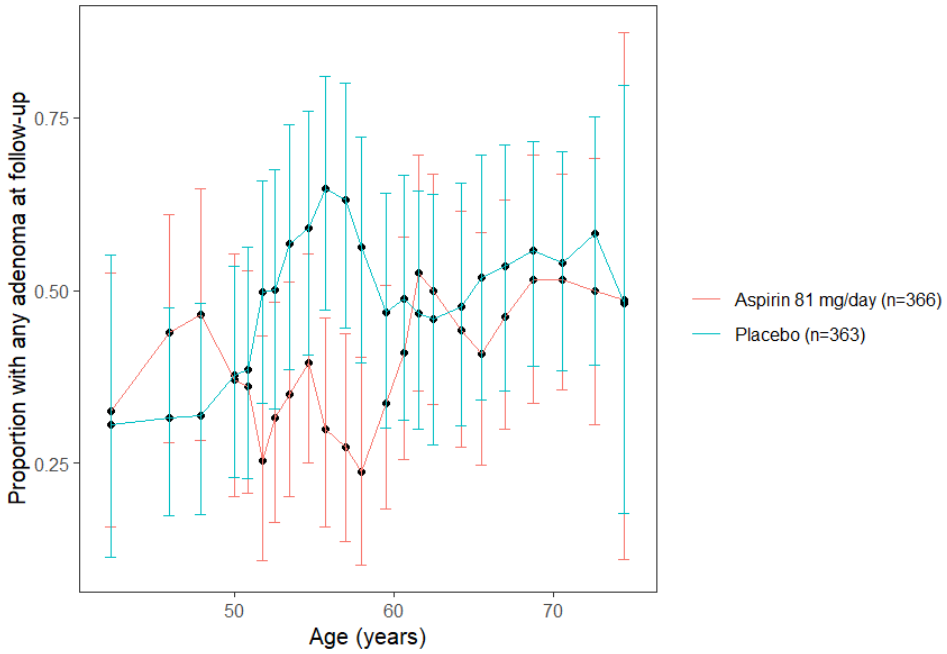


Figure 5.15: B-STEPP plot with general pattern of populations for the Aspirin/Folate Polyp Prevention Data. 81 mg/day (blue) vs. placebo groups (red), with 90% credible intervals of the marginal posterior distributions. *Arithmetic mean* assumption to model dependence.

As observed, the plot closely resembles the one obtained above but provides a more granular representation of the subset estimates. Further research is needed to investigate

under which conditions increasing the number of overlapping subpopulations enhances granularity without compromising the precision of the estimates.

5.6 Discussion and Future directions

In this work we have developed a Bayesian version of STEPP for both binary and survival data. Our approach has been to define the general structure, and show applications starting with base models that potentially cover various data behaviors. In particular, we implemented a hierarchical binary model for binary outcomes and a hierarchical exponential model and a lognormal one for survival outcomes. Finally, we have shown how these methods can be applied to realistic data. The idea is to extend in future work B-STEPP implementing other survival models based on different distributions such as loglogistic, Gamma, Gompertz, Gen Gamma, GenF, and include them in an appropriate R package. Once all these models have been constructed one can choose the best one in concrete applications relying on model fit statistics (such as AIC/BIC/WAIC).

There are different interesting extensions of the model above. The one that seems the most promising and interesting is related to a different (Bayesian) way of constructing the overlapping subpopulations. The first choice we have to make is to choose if we want to specify a priori the number of overlapping subpopulations, or to let the model suggest that. If we decide to specify a priori the number of intervals m we can think of a Bayesian model that works as follows:

1. specify as in the sliding window approach, the values of r_1 and r_2;

2. sample m disjoint sets following the model below, while paying attention to the value of r_2;

3. determine around the extremes of the intervals we sampled in the previous point the region of overlapping, so that r_1 and r_2 are achieved.

To introduce the model related to point 2, we define $\{l_0, ..., l_m\}$ the extremes of the disjoint intervals; naturally, $l_0 = \tilde{X}_{\min}$ and $l_m = \tilde{X}_{\max}$. Moreover, each $a_i \in (a_i, b_i)$ for $i = 1, ..., m-1$ where (a_1, b_i) is an open interval constructed so that the relation related to r_2 can be satisfied. Then the model can be constructed as follows:

$$l_i \sim U(a_i, b_i) \quad \text{for } i = 1, ..., m-1$$

$$Y_i|\eta_{l_j} \sim F(\eta_{l_j}) \quad \text{for } j = 1, ..., m$$

$$\eta_{l_j} \sim P(\alpha_{l_j}) \quad \text{for } j = 1, ..., m$$

where again $\boldsymbol{\eta}$ is explicitly related to $\tilde{\boldsymbol{\eta}}$ with a deterministic relation.

Once we have sampled the disjoint sets, point 3 follows, and we can obtain the final estimate as described in the previous sections.

Another way one can consider to determine the extremes of the overlapping intervals may be by maximizing the following quantity:

$$S(\tilde{\boldsymbol{\eta}}_{\boldsymbol{S}}) = \sum_{i=1}^{m-1} \mid \tilde{\eta}_i - \tilde{\eta}_{i+1} \mid$$

where $\tilde{\boldsymbol{\eta}}$ is estimated as described in the previous section, and with the maximization performed with different machine learning techniques (Grid search, Random search, Bayesian optimization,..). Also in this case, we have again to define some constraints on the intervals so that the conditions related r_1 and r_2 are satisfied.

Finally, another interesting extension, perhaps the most interesting one, involves finding a way to make the model suggest both the optimal number of subpopulations and their locations independently. In this case, a loss function that penalizes for a large number of intervals should be defined, and again, one should pay attention to the number of individuals in each subpopulation to avoid losing precision in the estimates.

References

Baron, J. A., Cole, B. F., Sandler, R. S., Haile, R. W., Ahnen, D., Bresalier, R., McKeown-Eyssen, G., Summers, R. W., Rothstein, R., Burke, C. A., et al. (2003). A randomized trial of aspirin to prevent colorectal adenomas. *Obstetrical & gynecological survey*, 58(8):538–539.

Bonetti, M. and Gelber, R. D. (2000). A graphical method to assess treatment–covariate interactions using the cox model on subsets of the data. *Statistics in medicine*, 19(19):2595–2609.

Bonetti, M. and Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–481.

Bonetti, M., Zahrieh, D., Cole, B. F., and Gelber, R. D. (2009). A small sample study of the stepp approach to assessing treatment–covariate interactions in survival data. *Statistics in medicine*, 28(8):1255–1268.

Coyle, D., Buxton, M. J., and O'Brien, B. J. (2003). Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. *Health economics*, 12(5):421–427.

Espinoza, M. A., Manca, A., Claxton, K., and Sculpher, M. J. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making*, 34(8):951–964.

Fisher, B., Dignam, J., Emir, B., Bryant, J., DeCillis, A., Wolmark, N., Wickerham, D. L., Dimitrov, N. V., Abramson, N., Atkins, J. N., et al. (1997). Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer. *Journal of the National Cancer Institute*, 89(22):1673–1682.

Glynn, D., Giardina, J., Hatamyar, J., Pandya, A., Soares, M., and Kreif, N. (2024). Integrating decision modeling and machine learning to inform treatment stratification. *Health Economics*, 33(8):1772–1792.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68.

Hu, L., Ji, J., and Li, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in medicine*, 40(21):4691–4713.

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of internal medicine*, 172(1):35–45.

Lagakos, S. W. et al. (2006). The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, 354(16):1667.

Lazar, A. A., Bonetti, M., Cole, B. F., Yip, W.-k., and Gelber, R. D. (2016). Identifying treatment effect heterogeneity in clinical trials using subpopulations of events: Stepp. *Clinical Trials*, 13(2):169–179.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194.

Wang, S.-J., O'Neill, R. T., and Hung, H. J. (2010). Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. *Clinical Trials*, 7(5):525–536.

Willke, R. J., Zheng, Z., Subedi, P., Althin, R., and Mullins, C. D. (2012). From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC medical research methodology*, 12:1–12.

Yip, W.-K., Bonetti, M., Cole, B. F., Barcella, W., Wang, X. V., Lazar, A., and Gelber, R. D. (2016). Subpopulation treatment effect pattern plot (stepp) analysis for continuous, binary, and count outcomes. *Clinical Trials*, 13(4):382–390.

Chapter 6

Discussion

6.1 Discussion

The aim of this work has been to introduce novel methodologies for health economics analysis in non-standard contexts.

In Chapter 2, we introduced Inverse Target Trial Emulation (ITTE), a novel methodology designed to generate observational data starting from a preliminary randomized controlled trial (RCT) dataset. ITTE works by "inverting" all the relevant steps of a standard Target Trial Emulation (TTE) procedure. The first step of ITTE is related to learn the dependence structure of the original trial data. In the specific implementation of ITTE in Chapter 2, we used a Bayesian nested regression model to pursue this objective. In the second step, one has to simulate new randomized data and, in the final step of the procedure, systematically introduce different forms of bias to produce coherent and realistic observational data. We demonstrated that ITTE successfully reproduces various forms of selection, confounding, and time-related bias. In particular, we reproduced data with unbalanced populations, which can be an effect of both confounding and/or selection bias, with different methods, each one assuming a different degree of prior knowledge on how the unbalance originated. Moreover, we reproduced the immortal time bias, a time-related bias that arises from a misalignment of the start of follow-up with treatment assignment, resulting in a wrong definition of time zero. We then applied this methodology to assess the robustness of different methods commonly employed to adjust for bias within observational data, by obtaining unbiased estimates of relevant treatment efficacy measures and comparing them with those of the original trial.

ITTE has the advantage of being a suitable candidate for a unifying methodology that can integrate all the different methods for generating observational data and testing the robustness of causal inference methods. This is because even if the methodology was born to simulate observational data starting from initial RCT data, one can also simulate the starting trial data and obtain a generalized methodology. Having initial RCT data

helps in reproducing more coherent and realistic observational data, resulting in a more consistent methodology to evaluate causal inference methods.

In Chapter 2, we presented a possible implementation of ITTE with a specific way of learning the relevant dependence structure from the RCT data. In particular, as said, we employed a parametric Bayesian nested regression method as this implementation allows for better control of relevant variable distributions and a higher reproducibility. Despite this, assuming a parametric model to learn the outcome distribution translates to simulating observational data with the same distribution we assumed in the Likelihood of the Bayesian model. This leads to estimating treatment effects by potentially assuming the same outcome model used to simulate data, and initially, to learn the relevant dependencies from RCT data. For this reason, further investigation can be based on exploring the effectiveness of nonparametric methods for learning the relevant dependencies within RCT data in the first ITTE step. In particular, we will focus our attention on nonparametric copulas, nested Bayesian Additive Regression Trees (BART), and autoregressive models. Even if it is known that these types of methods have drawbacks too (e.g., necessity of a high sample size), this seems a promising direction to explore, to further prove the potential of ITTE in incorporating different methods that work under different hypotheses.

In Chapter 3, we introduced a novel methodology for computing the Expected Value of Sample Information (EVSI) in the presence of missing data. A crucial ingredient we defined in this novel methodology is a novel data-generating process within the EVSI computation process, which enables the simulation of data affected by different types of missingness. Specifically, we successfully reproduced both missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) missing mechanisms. After that, we used the multiple imputation method to impute missing observations and to obtain an unbiased estimate of the T statistic in the nonparametric regression method to compute EVSI. Finally, we demonstrated that the EVSI decreases

due to the additional layers of uncertainty associated with the missing mechanism and the multiple imputation method. Therefore, we proposed a computationally efficient method to determine the number of individuals required to match the information from an RCT with a specific sample size, thereby recovering the EVSI value of data with no induced missingness.

This methodology has been proven to be particularly useful in understanding the value of additional realistic evidence. Given the promising results presented in the chapter, it would be valuable to explore theoretical properties that could help better define, identify, and quantify the additional sources of uncertainty contributing to the decrease in EVSI. In this context, starting from a standard conjugate Bayesian model, we plan to compute explicitly the variance associated with the additional layers of uncertainty. Then, we would like to estimate the effects of those layers separately, also in less standard scenarios.

A possible limitation of the new methodology is related to the MNAR missing mechanism. It is known that multiple imputation can lead to biased estimates when the underlying missing mechanism is MNAR, resulting in a biased estimate of the statistic T in the nonparametric method for computing EVSI. This does not directly introduce an additional layer of uncertainty in the same way we refer to uncertainty in the standard context; however, it still impacts the estimation of EVSI. Therefore, it is valuable to further explore this scenario to better understand how a biased estimate of statistic T, and in general, of the relevant parameters, influences EVSI estimates. This becomes even more important as, in this case, the method for calculating the necessary sample size to match the information from an RCT may fail as it assumes the estimate of T to be unbiased.

In Chapter 4, we introduced a methodology for computing EVSI when evidence is collected from observational studies. In this chapter, we combine the general methodology to compute EVSI in non-standard contexts introduced in Chapter 3 with the ITTE methodology to generate observational data introduced in Chapter 2. We apply ITTE to

simulate observational data in the data-generating mechanism within the EVSI computation process. Following the general methodology, after generating observational data, we applied inverse probability weighting (IPW) to compute an unbiased estimate of the T statistic in the nonparametric regression method, which is used to calculate EVSI. We show that EVSI decreases with respect to the EVSI computed on the related trial (with no bias induced by ITTE). This time, the decrease in EVSI is related to the uncertainty induced by both ITTE and IPW. Moreover, we show that the greater the bias we induce, the more the EVSI decreases. Finally, we apply the method designed in Chapter 3 to determine the number of individuals required to match the information from an RCT with a specific sample size, this time recovering the EVSI value of data with no induced bias.

This methodology allows combining the high availability and feasibility of observational data with such a powerful instrument for decision-making and research prioritization as EVSI, opening up different potential applications in various real-life scenarios in HTA. In a similar manner to Chapter 3, in subsequent work, one may conduct a further investigation into the theoretical properties that would allow us to define, identify, and quantify the different layers of uncertainty that cause the decrease in the EVSI value. We plan to start with Bayesian conjugate models, along with more standard methods of inducing bias. We will then move to the analysis and estimation of the different layers in more complex scenarios. Moreover, as we observed that the EVSI computation in this scenario is much more sensitive to the confounders we simulate in the original RCT data, we will also investigate this property further to understand how to properly account for this behavior in complex scenarios. This last property is, at the current time, the main limitation of this methodology, as it risks requiring researchers to perform computationally intensive calculations to obtain robust estimates.

Finally, in Chapter 5, we introduced a Bayesian approach to investigate heterogeneity of treatment effect (HTE) in two-treatment experimental studies. We started with

Subpopulation Treatment Effect Pattern Plot (STEPP), a methodology for investigating HTE that involves dividing the populations into overlapping subgroups and analyzing the treatment effects of these subgroups by properly modeling the dependence between subgroups. From this, we defined a Bayesian version of STEPP (B-STEPP), modeling relevant dependencies in a flexible way. We applied the novel methodology to analyze two real studies. The first refers to a binary outcome (see §5.3.1), while the second involves a survival setting (see §5.3.2), where we implemented both an exponential model and a lognormal model for the outcome. We demonstrated that B-STEPP successfully models dependence in a flexible way and exhibits better performance compared to standard subgroup analysis.

As STEPP, B-STEPP also reduces the drawbacks associated with standard subgroup analysis by working with overlapping subgroups, rather than disjoint ones. Moreover, preliminary analysis on the computation performance of B-STEPP shows promising results with respect to frequentist STEPP. The main current limitation of B-STEPP that opens the most interesting direction to explore is related to its parametric implementation, as specifying the correct outcome model is important to obtain good estimates. This is the case for the survival outcome, as the method implemented is equivalent to the nonparametric implementation for a binary outcome. For this reason, the next step will be to implement a Bayesian nonparametric version of B-STEPP. Other potential extensions discussed in §5.6 relate to implementing a general version of B-STEPP, in which the model would suggest both the optimal number of subgroups and their locations by itself.

Although B-STEPP and the Value of Information (VoI) methods proposed in this work, originate from different research lines, they can be naturally connected in applications where understanding and quantifying heterogeneity is key. For example, subgroup analysis has already been used to assess the value of heterogeneity through the implementation of VoI methods for subgroup analyses (Espinoza et al. (2014)). In this context, B-STEPP can identify patterns of treatment effect variation across patient characteristics,

while the EVSI methods developed in this work can quantify the value of collecting additional, potentially observational, data, accounting for missingness and confounding, to further characterize that heterogeneity. Similarly, when heterogeneity is explored within microsimulation models, B-STEPP and ITTE could be jointly applied to generate and analyze synthetic patient-level data, linking subgroup exploration with the economic value of reducing uncertainty.

The methodological work developed in this thesis can be used for various realistic applications different from the ones highlighted in this work. For example, it provides the foundation for extending EVSI to settings where future evidence is expected to come from observational data, as in Access with Evidence Development schemes. In a similar framework as the one highlighted in Claxton et al. (2016), AEDs allow conditional approval of promising therapies while planning the prospective collection of real-world evidence to support a later reassessment. In this context, EVSI must reflect not only the uncertainty in clinical outcomes, but also the characteristics of real-world data, such as missingness and confounding, that arise once treatments are used in routine practice. In future work, we will apply this methodology to a concrete case study: bevacizumab for metastatic colorectal cancer. To compute the EVSI, we will simulate the range of patient-level outcomes that could have been collected in Ontario had an AED been implemented in 2006, using only the evidence available at that time. The real-world data from the ON mCRC cohort will then be used solely to validate the realism of the simulated observational datasets. This application will illustrate how extending EVSI to observational data can support AED decision-making by quantifying the value of evidence expected to be generated through routine clinical practice following provisional approval.

Finally, it can be interesting to also discuss the practical feasibility of some of the different novel methodologies introduced in this work. In particular, related to the first three chapters, even if it is true that individual-level RCT data with coherent PICO structure may not always be fully available; in many applied contexts sufficient information

can still be reconstructed from multiple sources, such as published RCTs, meta-analyses, routinely collected real-world datasets, disease registries, and experts opinions', to inform the relevant parameters of the model. Even when some parameters are uncertain or only partially informed, these methods can still be applied by specifying plausible ranges and conducting sensitivity analyses. In this sense, while perfect RCT-level information is not always accessible, the combination of heterogeneous evidence sources typically provides enough structure to make the implementation of the introduced methodologies practically feasible in many contexts.

Overall, this thesis has contributed to the field of statistical methods for health economics both by defining a suitable candidate for a unified approach to test causal inference methods and simulate observational data (ITTE), and by extending the use of powerful statistical techniques to more realistic contexts (realistic EVSI) and with more flexible approaches (B-STEPP). All the novel methodologies have been applied to real data or realistic contexts to show their potential in addressing real health-related methodological problems.

References

Claxton, K., Palmer, S., Longworth, L., Bojke, L., Griffin, S., Soares, M., Spackman, E., and Rothery, C. (2016). A comprehensive algorithm for approval of health technologies with, without, or only in research: the key principles for informing coverage decisions. *Value in Health*, 19(6):885–891.

Espinoza, M. A., Manca, A., Claxton, K., and Sculpher, M. J. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making*, 34(8):951–964.

Chapter 7

Appendix

7.1 Appendix

We now list the relevant distributions and parameters we assumed to perform the previous simulations. We start by reporting the assumptions related with the analysis of the *help data* §2.3.1 dataset to testing the robustness of the listed causal inference methods.

Variable	Distribution	Mean	Variance	Truncation [left,right]
	Likelihood			
$dayslink_i$	$\text{TN}(\mu_i, \sigma_1^2)$	μ_i	σ_1^2	$[0, \infty)$
	$\mu_i = \theta_{90} + \theta_{91} \cdot treat_i + \ldots + \theta_{99} \cdot drugrisk_i$			
$gender_i$	$\text{Binomial}(1, \theta_0)$	–	–	–
$daysdrink_i$	$\text{TN}(\theta_2, \sigma_2^2)$	θ_2	σ_2^2	$[0, \infty)$
$daysanysub_i$	$\text{TN}(\theta_3, \sigma_3^2)$	θ_3	σ_3^2	$[0, \infty)$
pcs_i	$\text{TN}(\theta_4, \sigma_4^2)$	θ_4	σ_4^2	$[14, 75]$
mcs_i	$\text{TN}(\theta_5, \sigma_5^2)$	θ_5	σ_5^2	$[6.6, 62]$
$cesd_i$	$\text{Binomial}(60, \theta_6)$	$60\theta_6$	$60\theta_6(1-\theta_6)$	–
$sexrisk_i$	$\text{Binomial}(21, \theta_7)$	$21\theta_7$	$21\theta_7(1-\theta_7)$	–
$drugrisk_i$	$\text{Binomial}(21, \theta_8)$	$21\theta_8$	$21\theta_8(1-\theta_8)$	–
	Priors			
$\theta_0, \theta_6, \theta_7, \theta_8$	$\text{Unif}(0, 1)$	–	–	$[0, 1]$
θ_2, θ_3	$\mathcal{N}(100, 100)$	100	100	–
θ_4	$\text{TN}(45, 100)$	45	100	$[0, \infty)$
θ_5	$\text{TN}(34, 100)$	34	100	$[0, \infty)$
θ_{90}	$\mathcal{N}(0, 100)$	0	100	–
θ_{91}	$\mathcal{N}(-100, 100)$	-100	100	–
θ_{92}-θ_{99}	$\mathcal{N}(0, 100)$	0	100	–
σ_1^2	$\text{TN}(5, 2)$	5	2	$[0, \infty)$
σ_2^2-σ_5^2	$\text{TN}(10, 10)$	10	10	$[0, \infty)$

Table 7.1: Likelihood and prior distributions for model in §2.2.2.1 applied to the *help dataset*.

We now list the parameters and relevant distributions for the survival study with GBSG2 data.

Variable	Distribution	Mean	Variance	Truncation [left,right]
	Likelihood			
t_i	Exponential(λ_i)	$1/\lambda_i$	$1/\lambda_i^2$	$[0, \infty)$
	$\lambda_i = \exp(\beta_0 + \beta_1 \cdot \text{treat}_i + \beta_2 \cdot \text{tgrade}_i + \beta_3 \cdot \text{pnodes}_i)$			
age_i	TN(θ_0, σ_0^2)	θ_0	σ_0^2	$[20, \infty)$
$menostat_i$	Bernoulli($p[i]$)	$p[i]$	–	–
$tsize_i$	TN(θ_2, σ_2^2)	θ_2	σ_2^2	$[0, \infty)$
$tgrade_i$	Binomiale($3, \theta_3$)	$3 \cdot \theta_3$	$3 \cdot \theta_3(1 - \theta_3)$	–
$pnodes_i$	TN(μ_i, σ_4^2)	μ_i	σ_4^2	$[0, \infty)$
$progrec_i$	TN(θ_5, σ_5^2)	θ_5	σ_5^2	$[0, \infty)$
$estrec_i$	TN(θ_6, σ_6^2)	θ_6	σ_6^2	$[0, \infty)$
	Priors			
$\beta_0 - \beta_3$	$\mathcal{N}(0, 25)$	0	25	–
θ_0	$\mathcal{N}(50, 100)$	50	100	–
θ_{10}	$\mathcal{N}(0, 100)$	0	100	–
θ_{11}	$\mathcal{N}(0, 100)$	0	100	–
θ_2	TN($50, 100$)	50	100	$[0, \infty)$
θ_3	Unif($0, 1$)	–	–	$[0, 1]$
θ_{40}	TN($5, 10$)	5	10	$[0, \infty)$
θ_{41}	$\mathcal{N}(0, 10)$	0	10	–
θ_5	TN($100, 100$)	100	100	$[0, \infty)$
θ_6	TN($100, 100$)	100	100	$[0, \infty)$
σ_0^2	TN($5, 2$)	5	2	$[0, \infty)$
$\sigma_2^2 - \sigma_6^2$	TN($10, 10$)	10	10	$[0, \infty)$

Table 7.2: Likelihood and prior distributions for model in §2.2.2.1 applied to the *GBSG2* dataset.