UNIVERSITÀ DEGLI STUDI FIRENZE

UNIVERSITÀ DEGLI STUDI DI PERUGIA

Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN STATISTICA
CICLO XXXII

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof.Graziano Gentili

Causal inference in irregular designs with intermediate variables and multivariate outcomes and the external validity of RCT results

Settore Scientifico Disciplinare SECS-S/01

Dottoranda
Federica Licari

Tutore
Prof.ssa Alessandra Mattei

Coordinatore
Prof.Graziano Gentili

Anni 2016/2019

Dedicated to my Sister
the best Cata I know
I love you

Contents

Introduction

My PhD thesis is organized in three main chapters which are the realization of three main works. The first work reported in Chapter 1 is a joint work with my tutor Alessandra Mattei, associate professor in the Department of Statistics, Computer Science, Applications "G. Parenti" (DiSIA) of Florence, where I started my PhD in 2016. We titled our work "Assessing causal effects of extra compulsory learning on college students' academic performances". The second work reported in Chapter 2 is a joint work with Alessandra Mattei and the associate professor in DiSIA Michela Baccini. We titled our work "The role of prior knowledge and numerical information in communicating epidemiological results to the population. A Bayesian Principal Stratification approach". Finally, the third work, reported in Chapter 3, is a joint work with Corwin Zigler, associate professor in the Department of Statistics and Data Sciences (College of Natural Sciences) and in the Department of Women's Health (Dell Medical School) of Austin, and with Joseph Koopmeiners, associate professor and interim division head in biostatistics in School of Public Health in Minnesota. We titled this work "Stratification via Propensity Score and Bayesian Additive Regression Trees approaches to estimate PATT from RCT".
In these three works we formulate causal research questions rather than association research question. Standard statistical analyses aim to assess the associations between variables, conducting inference to estimate parameters which govern the joint distributions of the variables involved. The main goal of standard statistical analyses is to learn from informations of the past in order to better predict the future according with new informations provided. The assumption behind the standard statistical analyses is that the experimental conditions do not change between past and future.
Causal analyses work under the assumption that experimental conditions may change because of several factors, e.g. changes due to the treatments or external interventions. The term "treatment" is referred to any action applied to a unit, e.g. physical object, individual person. A unit changes during the time, therefore, from a causal point of view one unit in a specific time is different from the same in an other time. Thus, a given unit is different by itself after the application of a treatment. Let apply a treatment to a unit and let observe the effect of this treatment on the outcome for that unit. Since that unit after the application of the treatment is not the same of those one it was before the application, we are not able to observe what would have been the outcome if an other type of treatment had been applied.
What the effect of a treatment on a given unit is rather than the effect of an other

treatment on the same unit, defines a causal research question.
Specifically, given a treatment X the potential outcome for that treatment of a given unit is the outcome we would observe if treatment X would be applied to that unit. Given two different treatments X and Y the comparison between the related potential outcomes of a given unit is defined as causal effect.
Unfortunately, for every unit we almost observe one potential outcome, since that at a fixed time we can apply only one treatment and after the application the unit is not the same of those one before the application.
Thus, we cannot learn the value of causal effect from just a single realized potential outcome.
If we want to estimate the causal effect we need to compare observed outcomes, that is observed realizations of potential outcomes considering multiple units.
Involving more units requires also to make assumptions related to the assignment mechanism of the several treatments involved. One of the aim of these assumptions is to guarantee the absence of manipulation of the assignment of the treatment.
The randomization ensures that the assignment of the treatment is independent of the potential outcomes as well as independent of other variables. Therefore, the missing potential outcomes are missing at random. Thus, comparisons between realizations of potential outcomes under a given treatment and those under an other treatment produce unbiased estimate of the causal effect of interest.
In the three works reported in my PhD thesis, we address the causal research questions within the potential outcome framework highlighting the key role of the randomization. Specifically, for the first work, we adopt the probabilistic formulation of the sharp-Regression Discontinuity Designs, introduced by Li et al. (2015), to assess the causal effect of the extra compulsory learning on college students' academic performances dealing with multivariate outcome.
In the second work randomization is given by design. For these two works we also apply the principal stratification approach to conduct causal inference. Principal stratification approach consists in dividing units in latent strata defined by the vector of the all possible realizations of potential outcomes of a fixed post treatment variable, or more than one post treatment variable. The basic idea behind principal stratification approach is that comparisons of potential outcomes within each stratum are always causal effects. In the two works we appeal to principal stratification for different reasons. In the first work, we use the principal stratification approach in order to deal with the problem of truncation by "death". In the second work, we appeal to the principal stratification approach in order to investigate the difference of the treatments' effects taking into account the presence of an intermediate variable which may be affected by the treatments itself.
In the third work we move the focus from the randomization of the assignment mechanism of treatment versus the randomization of the sample selection mechanism, dealing with the problem of the external validity of Randomized controlled trial results.
Briefly, below it is reported an abstract of each work.

Assessing causal effects of extra compulsory learning on college students' academic performances

In this work we investigate about the role of the entry test in an Italian University. In Italian universities candidates freshmen must take an entrance exam. Candidates who obtain a test score less than or equal to a prefixed threshold may enrol in the University but must comply with additional educational requirements, named OFA (Italian acronym for Obblighi Formativi Aggiuntivi). The OFA assignment rule appeals to a (sharp) Regression Discontinuity (RD) design with the entrance exam score acting as forcing variable. In this work we aim at assessing causal effects of the OFA status using data from a School of Engineering of a specific Italian State University. For sub-populations of units for which our RD design can be described as a local randomized experiment, we draw inference on the causal effects of OFA on students' academic performances measured using two variables: the number of University credits awarded and the corresponding average grade. These outcome variables suffer from the problem of truncation by death. The truncation is due to the academic career status: after the entry test participants may decide to not enrol or if they enrol they could not take exams. We deal with these issues using the framework of principal stratification and adopting a Bayesian approach to inference. Our analyses reveal weak positive OFA effect on University credits and negligible effects on average grade for a suitable selected sub-population of students who enrol in University and take exams.

The role of prior knowledge and numerical information in communicating epidemiological results to the population. A Bayesian Principal Stratification approach

In this work we contribute to the existing literature on the role of the format used on communicating results in the field of epidemiology. We use data from a recent randomized experiment conducted on a sample of residents in Livorno, an Italian city classified as a high environmental risk area according to the Seveso Directive. The main goal is to understand if people judgements about local risks for population health are influenced by the statistical indicators used in risk communication (Baccini et al., 2019). The enrolled people are randomized to reply to different questionnaires, which define our binary treatment, where the same epidemiological results are expressed using two types of risk indexes: % of excess risk (%ER) and time needed to harm (TNH). Then the respondents are requested: 1) to classify the mortality results for three different cancers (Sexual Gland Cancer (SGC), Tthyroid Gland Cancer (TGC) and Lung Cancer (LC)), according to their concern: from most alarming to least alarming. 2) Express in a scale from 0 to 10 their concern about mortality from cancer in Livorno area (Y_D). We are particularly interested in investigating the difference of the treatments' effects with respect to the contribution, in responding to the questionnaire items, of the a priori knowledge about environmental and health issues, which can be itself affected by the treatment. We deal with this issue using a principal stratification approach. Formally,

we define principal strata on a post-treatment variable denoting if people declare to have answered basing of the communicated numbers or on other elements. In particular, we focus on the relative effect of the two risk indexes in the stratum of people that base their answer on numbers under both treatments, and we evaluate if the use of an indicator rather than the other inhibits the respondent from basing her/his answer on the numerical results. We adopt a model based Bayesian approach to inference.
Results differ among the outcomes analysed, suggesting a sensible difference between treatments' effects on classification SGC as most alarming, as well as for TGC and outcome Y_D; on the other hand negligible effect on classification of LC.

Stratification via Propensity Score and BART approaches to estimate PATT from RCT

In this work we contribute to the existing literature on the external validity of Randomized Controlled Trials (RCTs). The aim of RCTs is to test the effectiveness of a new treatment. The random assignment of the treatment in these types of experimental designs ensures the absence of manipulation of the results. However, results obtained by RCTs may not have external validity in such case, that is they are not generalizable to a Target Population (TP). Several factors may cause the absence of generalizability of the results from RCT, e.g. the eligibility criteria for the inclusion in the RCT provide a sample not well representative of the TP and/or the treatment may have different version between RCT sample and TP.
Here, we formulate estimators based on two approaches: 1)the Stratification via Propensity Score (SvPS) approach, where the propensity score in this case is the probability to be selected in the RCT sample; 2)Bayesian Additive Regression Trees (BART) approach. We use these estimators for double aims: evaluate the external validity of RCT results and, in case for which RCT results are generalizable, estimate the Population Average Treatment effect on Treated (PATT) from the Sample Average Treatment Effect (SATE) within the potential outcomes framework.
The estimators we propose provide estimates of the PATT with negligible bias. Performance of the SvPS estimator does not differ from those one of BART estimator.
Finally, we present an analysis of Medicare simulated data where we first evaluate the external validity of a Corevalve trial results, then we estimate the PATT effect of the Trans-catheter Aortic Valve Replacement (TAVR) vs the Surgical Aortic Valve Replacement (SAVR) on days until the first complication occurrence.

Chapter 1

Assessing causal effects of extra compulsory learning on college students' academic performances

KEYWORDS: Bayesian inference, Local randomization, Potential outcomes, Multiple Outcomes, Principal Stratification, Regression discontinuity designs

1.1 Introduction

In Italy, students need to take an entry test before enrolling in University. The aim of the pre-test is to give a measure of students' aptitude in various areas such as science, maths, and language skills. The test may provide useful information to both individuals and institutions. It allows candidate freshmen to self-evaluate their aptitude for the chosen college program. Schools can use the test results as a proxy of the skills of potential freshmen for subjects of their college programs and for efficiently planning teaching support for weaker students, who get a low score in the test. Recently, schools in Italian Universities have introduced so-called additional learning obligations, named OFA (Italian acronym for Obblighi Formativi Aggiuntivi): a participant with a test score less than or equal to a prefixed threshold may enrol in the University but s/he must comply with additional compulsory learning, by taking an extra exam.
In this work we are interested in assessing causal effects of OFA on students' career status and academic performances at the end of the first academic year for students who participated in the entrance exam for enrolling in the School of Engineering of a given Italian State University in 2011. We focus only on one school of a specific University because different Universities use different thresholds for assigning students to OFA, and the arguments and the structure of the entry tests are different across different types of schools depending on the type of major [1]. Students' career status is measured by an indicator equal to one for active students, that is, students who take and pass at least one exam during the first academic year and zero for inactive students, who do not either take or pass any exam. Academic performances are measured using the awarded number of University educational credits (UCs) and the corresponding average grade, defined

[1] We generally refer to the School of Engineering of a given Italian State University because, for confidentiality issues, we cannot name the specific institution.

as the weighted average of grades with weights given by the UCs. In Italy, a pre-fixed number of UCs is assigned to each university exam. One credit usually corresponds to 25 hours of work, including both class work (lessons, exercises, labs etc), as well as home study. For each academic year, full-time students are required to complete an average quantity of work sets at 60 credits. Credits are obtained by passing exams, so they are a measure of the quantity of a student's work. The average grade is a measure of the quality of the student's performance. The choice of focussing on these outcome variables is compelling. In fact, in Italy, a high percentage of students, around 20%, is classified as *inactive* at the end of the first academic year. Moreover, active students with a poor aptitude for the subjects of the chosen college program may not perform very well, taking a low number of exams with respect to planned number of exams and/or receiving low grades (see, e.g., the last report of the National Agency for the Evaluation of Universities and Research Institutes, ANVUR[2]). Therefore, assessing causal effects of OFA on these variables is of great interest in policy, providing evidence on whether OFA are effective in hindering students' inactivity and low academic performances.

The receipt of OFA depends on the test score falling below or above a predetermined threshold. This allocation rule motivates us to adopt the regression discontinuity (RD) design framework for evaluation. The RD design – a quasi-experimental design for causal inference – was first introduced in psychology by Thistlethwaite and Campbell (1960) and has became increasingly popular since the late 1990s in statistics, social science, economics and, more recently also in epidemiology and the medical sciences. In RD designs, the assignment to the treatment is determined, at least partly, by the realized value of a variable, usually called the forcing or running variable, falling on either side of a prefixed threshold. The basic idea underlying a RD analysis is that one can compare units with very similar values for the forcing variable, but different levels of treatment, to draw causal inference of the treatment at the threshold.

The Italian university OFA allocation rule defines a sharp RD design with the test score acting as forcing variable. In the sharp RD designs, the original form of the design, the treatment status is assumed to be a deterministic step function of the forcing variable. All students who obtain a test score falling below a prefixed threshold receive OFA, and all students who obtain a test score falling above the threshold do not receive OFA.

In the traditional approach to RD designs, the forcing variable is viewed as a pre-treatment covariate and RD designs are usually described as quasi-experimental designs with a non-probabilistic assignment mechanism. The literature has dealt with the lack of overlap invoking smoothness assumptions for the relationship between the outcome and the forcing variable, such as continuity of conditional regression functions (or the conditional distribution functions) of the outcomes given the forcing variable. Under these smoothness assumptions, which imply randomization at the single threshold value (Battistin and Rettore, 2008), we can identify causal effects at the threshold, but inference needs to rely on some kind of extrapolation: observations near the known threshold are

[2]The full report is available in https://www.anvur.it/wp-content/uploads/2019/01/ANVUR-Completo-con-Link.pdf

used to derive estimates of treatment effects at the threshold, using global polynomial series estimators or local-polynomial (non-)parametric regression methods and their asymptotic proprieties. See, e.g., Lee (2008); Lee and Lemieux (2010); Imbens and Lemieux (2008); Athey and Imbens (2016) and the edited volume by Cattaneo and Escanciano (2016) for literature reviews on the classical approach to RD design. In real applications, large-sample approximations might be unreliable, especially if the sample size around the threshold is small, and exact inference might be preferable.

Building on the original idea by Thistlethwaite and Campbell (1960), RD designs have been often described as designs that lead to locally randomized experiments around the threshold (e.g., Lee, 2008; Lee and Lemieux, 2010). Expanding on this interpretation, a recent strand of the literature (e.g., Cattaneo et al., 2015; Li et al., 2015; Sales and Hansen, 2015; Mattei and Mealli, 2016) has introduced a formal and well-structured definition of the conditions under which RD designs can be formally described as local randomized experiments. In this work we embrace this new perspective, adopting the approach described in Li et al. (2015) and Mattei and Mealli (2016).

This approach embeds RD designs in a framework that is fully consistent with the potential outcome approach to causal inference (e.g., Rubin, 1974, 1978), providing a formal definition of the hypothetical experiment underlying RD designs, based on a formal description of the assignment mechanism, as a unit-exchangeable stochastic function of covariates and potential outcomes.

A distinguishing feature of this approach is to view the forcing variable as a random variable with a probability distribution, rather than as a pre-treatment variable, and to assume that there exist at least a sub-population of units around the threshold where a local overlap assumption holds, and where the forcing variable, and therefore the treatment status, can be seen as randomly assigned. Under this framework focus is on local causal effects, that is, causal effects for sub-populations of units where the local overlap and local randomization assumptions hold.

In our study, this approach leads to focus on local causal effects of OFA on students' academic career and academic performances at the end of the first year for sub-populations of test participants with realized values of the test score falling in a neighbourhood around the threshold.

In this framework our work makes interesting methodological contributions. Causal inference on students' academic career and academic performances, measured by the awarded number of UCs and the average grade is, however, not trivial due to post-assignment complications. We set up an approach for defining and drawing inference on causal effects of OFA on multiple outcomes, for which potential outcomes are defined only on specific sub-groups of units. First, some students may decide to not enrol in the School of Engineering (preferring to not enrol at all, enrol in another school, or enrol in the School of Engineering of another University). Academic career and performances in the School of Engineering of interest for students who do not enrol are not defined: they are *truncated by death/ not enrolment* (Zhang and Rubin, 2003; Rubin, 2006). Second, some students who enrol in the School of Engineering may be inactive. For this type of

students the average grade is again not defined: it is *truncated by death/inactivity*. We focus on describing and addressing these complications in our study using the framework of principal stratification (Frangakis and Rubin, 2002). Specifically we first cross-classify test participants with respect to the joint potential values of the enrolment status and of the academic career status under each OFA status. Principal stratification makes it clear that (*i*) causal effects on students' academic career and the awarded number of UCs are well defined only for test participants who would always enrol in the School of Engineering we are focusing on irrespective of whether they receive or do not receive OFA (*always-enrolled students*); (*ii*) causal effects on average grade are well defined only for always-enrolled test participants who would always take and pass exams irrespective of whether they receive or do not receive OFA (*always-active students*). We use a model-based Bayesian approach for inference which is particularly attractive in RD settings where the analysis may rely on a small sample size, because it not relies on asymptotic approximations. From a Bayesian perspective, all inferences are based on the posterior distributions of causal estimands, which are functions of potential outcomes. Moreover the Bayesian approach is particularly useful for accounting for uncertainties and for pooling information from the data in complex settings like RD designs, especially in the presence of post-assignment complications. We specify a joint model for the awarded number of UCs and the average grade, which allows us to naturally account for the correlation between the two outcomes.

1.2 The OFA Data Set

We use data coming from the cohort of students who took the entry test for the Engineering area in 2011. The data were collected thanks to the collaboration between one School of Engineering and CISIA[3].

In 2011 the entrance exam for the schools of Engineering consisted of 80 items subdivided in 4 areas: Math-level 1 (20 items), Science (20 items), Logic (15 items), Reading Comprehension (15 items), and Math-level 2 (10 items). The test score is calculated as follows: 1 point for each right answer, -0.25 points for each wrong answer, and 0 points for each no-answer, so the test score takes on values in the interval $[-20, 80]$. The test score acts as forcing variable in our study. Formally, consider a sample of N subjects who participated in the entrance exam in 2011. Let S_i denote the test score for participant i. The treatment status, that is, the OFA status, denoted by Z_i, is a deterministic function of S_i: $Z_i = \mathbb{I}\{S_i \geq s_0\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function and $s_0 = 16$ is the prefixed threshold by the School of Engineering considered in the study. Thus, if a participant has a value of the test score, S_i, falling above the threshold $s_0 = 16$, $Z_i = 1$ and s/he does not receives OFA, and $Z_i = 0$ and s/he does receive OFA otherwise. Given the support of the test score variable, setting the threshold at $s_0 = 16$ is equivalent to require that a test participant correctly answers to at least 16% of the test items to

[3]CISIA is a para-university company born in 2005. CISIA organizes and supervises the entrance exam for various schools of Engineering and, recently, also for other schools of Italian State Universities.

enrol without OFA.
Four variables were measured after each participant i was assigned OFA status, Z_i: E_i^{obs}, the enrolment status, equal to 1 for participants who enrol in the school of Engineering, and 0 otherwise; C_i^{obs}, the career status, equal to 1 for active students, that is, students who pass at least one exam during the first academic year, and 0 for inactive students, that is, students who pass no exam during the first academic year; $Y_{i,UC}^{obs}$, the number of UCs awarded at the end of the first academic year; and $Y_{i,G}$, the average grade at the end of the first academic year. Let $\mathbf{Y}_i^{obs} = (Y_{i,UC}^{obs}, Y_{i,G}^{obs})$. It is worth noting that the career status, C_i^{obs}, and the two variables measuring students' academic performances, $Y_{i,UC}^{obs}$ and $Y_{i,G}^{obs}$, are not defined for test participants who do not enrol: they are truncated by 'death.' Because 'death' is due to the decision of not enrolling in the school of Engineering, we refer to those variables as *truncated by not enrolment* for test participants who decide to not enrol. We set $C_i^{obs} = Y_{i,UC}^{obs} = Y_{i,G}^{obs} = *$ if participant i does not enrol in the school of Engineering ($E_i^{obs} = 0$), where $*$ is a non-real value. Moreover students who enrol ($E_i^{obs} = 1$) but are inactive ($C_i^{obs} = 0$) award zero UCs ($Y_{i,UC}^{obs} = 0$) and thus, their average grade is not defined, that is, it is truncated by 'death.' We refer to average grade for inactive students as *truncated by inactivity*, because 'death' is due to inactivity. Again we set $Y_{i,G}^{obs} = *$ for inactive students.
For each test participant, i, we also observed a vector of covariates, $\mathbf{X}_i$: gender, high school grade, high school type (categorical variable with 4 levels) high school year (a binary variable that indicates if a participant finished the high school the year before the entry test), an indicator for irregular high school career, and an indicator for living far-away from the campus.
Our focus is on assessing causal effects of the OFA status on academic career, and number of UCs awarded at the end of the first year after the enrolment and corresponding average grade properly accounting for the irregular OFA assignment rule, which defines a (sharp) RD design, and the problem of truncation by 'death'.

1.3 Overlap Assumption and Potential Outcomes

Following Li et al. (2015) and Mattei and Mealli (2016), we view the forcing variable, S, as a random variable with a probability distribution, and formulate the following assumption:

Assumption 1.1. (Local Overlap). Let $\mathcal{U}$ be the random sample (or population) of units in the study. There exists a subset of units, $\mathcal{U}_{s_0}$, such that for each $i \in \mathcal{U}_{s_0}$, $Pr(S_i < s_0) > \epsilon$ and $Pr(S_i \geq s_0) > \epsilon$,
for some sufficiently large $\epsilon > 0$.

Assumption 1.1 is essentially a local overlap assumption, implying that there exists at least one sub-population of subjects, each of whom has a probability of having a value of the forcing variable (test score) falling on both sides of the threshold, s_0, sufficiently faraway from both zero and one. It is worth noting that Assumption 1.1 does not require

that the sub-population $\mathcal{U}_{s_0}$ is unique; it only requires that there exists at least one sub-population $\mathcal{U}_{s_0}$ for which Assumption 1.1 holds. Assumption 1.1 is a local overlap assumption in the sense that a subject who obtain a value of the test score falling very faraway from the threshold does not probably belong to any sub-population $\mathcal{U}_{s_0}$ and may have a zero probability of having a value of the test score falling on the other side of the threshold.

Under Assumption 1.1 each subject belonging to a sub-population $\mathcal{U}_{s_0}$ has a non-zero marginal probability of being assigned to either receiving or not receiving OFA: $0 < Pr(Z_i = 1) < 1$ for all $i \in \mathcal{U}_{s_0}$.

Under Assumption 1.1 focus is on local causal effects for a sub-population, $\mathcal{U}_{s_0}$, rather than on causal effects at the threshold, which are the causal estimands typically considered in RD designs. We frame our discussion in the potential outcome approach to causal inference, where causal effects are defined as comparisons of potential outcomes on a common set of units. Therefore, to formally define causal effects we need to introduce potential outcomes for each post-treatment variable.

The test score is the assignment variable: each unit in the sub-population $\mathcal{U}_{s_0}$ can be exposed to alternative values of the test score. Therefore, in principle, potential outcomes need to be defined as function of the forcing variable. Let $\mathbf{s} = [s_i]_{i \in \mathcal{U}_{s_0}}$ be a vector of values of the test score for units belonging to a sub-population $\mathcal{U}_{s_0}$. For each unit $i \in \mathcal{U}_{s_0}$ let $E_i(\mathbf{s})$, $C_i(\mathbf{s})$, $\mathbf{Y}_i(\mathbf{s})$ denote the potential outcomes for the post-treatment variables, E_i, C_i and $\mathbf{Y}_i$, respectively. We restrict the set of potential outcomes by assuming a modified Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980), specific to RD settings (Mattei and Mealli, 2016):

Assumption 1.2. (Local RD-SUTVA). For each $i \in \mathcal{U}_{s_0}$, consider two treatment statuses $z_i^{'} = \mathbb{I}(s_i^{'} \leq s_0)$ and $z_i^{''} = \mathbb{I}(s_i^{''} \leq s_0)$, with possibly $s_i^{'} \neq s_i^{''}$. If $z_i^{'} = z_i^{''}$, that is, if either $s_i^{'} \leq s_0$ and $s_i^{''} \leq s_0$, or $s_i^{'} > s_0$ and $s_i^{''} > s_0$, then $E_i(\mathbf{s}^{'}) = E_i(\mathbf{s}^{''})$, $C_i(\mathbf{s}^{'}) = C_i(\mathbf{s}^{''})$, and $\mathbf{Y}_i(\mathbf{s}^{'}) = \mathbf{Y}_i(\mathbf{s}^{''})$.

Assumption 1.2 guarantees (1) the absence of interference between individuals, i.e., the value of the test score (and the OFA status) of one unit do not affect outcomes of other units; and (2) potential outcomes depend on the forcing variable solely through the treatment indicator, z, but not directly, so that, values of the test score variable leading to the same OFA status define the same potential outcomes.

Assumption 1.2 allows us to write $E_i(\mathbf{s})$, $C_i(\mathbf{s})$, and $\mathbf{Y}_i(\mathbf{s})$ as $E_i(z_i)$, $C_i(z_i)$ and $\mathbf{Y}_i(z_i)$, respectively, for each unit $i \in \mathcal{U}_{s_0}$, where $z_i = \mathbb{I}\{s_i \geq s_0\}$. Therefore under local RD-SUTVA for each unit within $\mathcal{U}_{s_0}$ there exist only two potential outcomes for each post-treatment variable, $\underline{E}_i \equiv (E_i(0), E_i(1))$, $\underline{C}_i \equiv (C_i(0), C_i(1))$ and $\underline{\mathbf{Y}}_i \equiv (\mathbf{Y}_i(0), \mathbf{Y}_i(1))$: they are the values of each post-treatment variable if the unit had a value of the test score falling below and above the threshold, respectively.

1.4 Principal Stratification and Principal Causal Effects

Under local overlap and local RD-SUTVA, causal effects are defined as comparisons of the no OFA versus OFA potential outcomes for a common set of units in $\mathcal{U}_{s_0}$. They are local causal effects in that they are causal effects for units belonging to a sub-population $\mathcal{U}_{s_0}$. We focus on finite population average causal effects. Let $N_{\mathcal{U}_{s_0}}$ be the number of students in $\mathcal{U}_{s_0}$: $N_{\mathcal{U}_{s_0}} = \sum_{i=1}^{N} \mathbb{I}\{i \in \mathcal{U}_{s_0}\}$.
The finite population average causal effect of being exempt from OFA versus not being exempt from OFA on the decision to enrol is:

$$ACE_{\mathcal{U}_{s_0}}(E) = \mathbb{E}\left[E_i(1) \mid i \in \mathcal{U}_{s_0}\right] - \mathbb{E}\left[E_i(0) \mid i \in \mathcal{U}_{s_0}\right] = \frac{1}{N_{\mathcal{U}_{s_0}}} \sum_{i \in \mathcal{U}_{s_0}} \left[E_i(1) - E_i(0)\right]. \tag{1.1}$$

The definition of causal effects on students' academic career and students' academic performances is challenging due to the problem of truncation by death, that is, by enrolment and inactivity: for $z = 0, 1$, $C_i(z) = *$, and $Y_{i,UC}(z) = *$ if $E_i(z) = 0$; and $Y_{i,G}(z) = *$ if either $E_i(z) = 0$ or $E_i(z) = 1$ and $C_i(z) = 0$.

1.4.1 Principal Strata

We deal with the problem of truncation by death using the principal stratification framework (Frangakis and Rubin, 2002). A principal stratification with respect to a post-treatment variable (which may be multivariate) defines a cross-classification of subjects into groups, named principal strata, defined by the joint potential values of that post-assignment variable under each of the assignments being compared.
In our study we classify participants in the entry test with respect to joint indicator for enrolment and academic career status. Formally, the principal strata are defined by the values of the following vector: $B_i = (\underline{E}_i, \underline{C}_i) = (E_i(0), E_i(1), C_i(0), C_i(1))$. Thus, there are nine principal strata representing nine types of individuals:

Never-enrolled students. Students who would not enrol irrespective of the OFA status:

$$NE_{**} \equiv \{i : E_i(0) = 0, E_i(1) = 0, C_i(0) = *, C_i(1) = *\}$$

Note that the academic career is truncated by not enrolment for never-enrolled test-participants.

OFA-compliant inactive students. Students who would enrol but would be inactive if exempt from OFA and would not enrol if not exempt from OFA:

$$OC_{*0} \equiv \{i : E_i(0) = 0, E_i(1) = 1, C_i(0) = *, C_i(1) = 0\}$$

OFA-compliant active students. Students who would enrol and would be active if

exempt from OFA and would not enrol if not exempt from OFA:

$$OC_{*1} \equiv \{i : E_i(0) = 0, E_i(1) = 1, C_i(0) = *, C_i(1) = 1\}$$

We refer to the union of OFA-compliant inactive and active students as OFA-Compliers:

$$OC = OC_{*0} \cup OC_{*1} \equiv \{i : E_i(0) = 0, E_i(1) = 1\}.$$

They are students who would enrol if exempt from OFA and would not enrol if not exempt from OFA.

Always-active students. Students who would always enrol and would always have an active academic career irrespective of the OFA status:

$$AE_{11} \equiv \{i : E_i(0) = 1, E_i(1) = 1, C_i(0) = 1, C_i(1) = 1\}$$

Never-active students. Students who would always enrol but would never have an active academic career irrespective of the OFA status:

$$AE_{00} \equiv \{i : E_i(0) = 1, E_i(1) = 1, C_i(0) = 0, C_i(1) = 0\}$$

OFA-inactive students. Students who would always enrol irrespective of the OFA status, but who would have an active academic career if exempt from OFA, and an inactive academic career if not exempt from OFA:

$$AE_{01} \equiv \{i : E_i(0) = 1, E_i(1) = 1, C_i(0) = 0, C_i(1) = 1\}$$

OFA-active students. Students who would always enrol irrespective of the OFA status, but who would have an inactive academic career if exempt from OFA, and an active academic career if not exempt from OFA:

$$AE_{10} \equiv \{i : E_i(0) = 1, E_i(1) = 1, C_i(0) = 1, C_i(1) = 0\}$$

We refer to the union of always-active, never-active, OFA-inactive, and OFA-active students as *always-enrolled students*:

$$AE = AE_{11} \cup AE_{00} \cup AE_{01} \cup AE_{10} =\equiv \{i : E_i(0) = 1, E_i(1) = 1\}.$$

They are students who would always enrol in the University irrespective of the OFA status.

OFA-defiant inactive students. Students who would enrol but would have an inactive academic career if not exempt from OFA and would not enrol if exempt from OFA:

$$OD_{0*} \equiv \{i : E_i(0) = 1, E_i(1) = 0, C_i(0) = 0, C_i(1) = *\}$$

OFA-defiant active students. Students who would not enrol if exempt from OFA and would enrol and would have an active academic career if not exempt from OFA:

$$OD_{1*} \equiv \{i : E_i(0) = 1, E_i(1) = 0, C_i(0) = 1, C_i(1) = *\}$$

We refer to the union of inactive and active defiant-enrolled students as *OFA defiers*:

$$OD = OD_{0*} \cup OD_{1*} = \equiv \{i : E_i(0) = 1, E_i(1) = 0\}.$$

They are students who would not enrol if exempt from OFA and enrol if not exempt from OFA.

In our study appears very plausible to rule out the existence of OFA defiers, invoking the following monotonicity assumption:

Assumption 1.3. (Monotonicity of Enrolment) For all $i \in \mathcal{U}_{s_0}$, $E_i(0) \leq E_i(1)$.

Underlying this assumption is the notion that test participants who do not receive OFA are unlikely to decide to not enrol whether they would have enrolled if they had received OFA. Monotonicity would be violated if, for instance, some test participants, who would not enrol when assigned to OFA, would decide to enrol when not assigned to OFA, because they feel that they could benefit from OFA. We believe that it is unlikely there are many students of this type. In fact, test participants who receive OFA and enrol must attend an additional course and take the corresponding exam; they can start their academic career by taking University exams according to their engineering program, only if they successfully pass this extra exam. Nevertheless, the assumption of monotonicity of enrolment is not testable. Therefore, we also conduct a Bayesian sensitivity analysis with respect to the assumption of monotonicity of enrolment. We focus on a simplified setting where focus is on assessing causal effects of OFA on academic career, C, and investigate how the posterior distributions of the proportions of principal strata defined by the joint potential value of the enrolment indicator under each OFA status, $\underline{E}_i = (E_i(0), E_i(1))$, and the posterior distribution of the causal effect on academic career derived under monotonicity change when we relax the monotonicity assumption. The sensitivity analysis suggests that results are robust with respect to the monotonicity assumption. Moreover we find that the posterior probability that there are OFA defiers is very small, less than 5% when the monotonicity assumption is relaxed (see Appendix 1.9.3 for technical details on the sensitivity analysis w.r.t. Assumption 3 and results from it). Given the insights underlying the monotonicity assumption and the results from the sensitivity analysis, we make Assumption 1.3 throughout our discussion.

Under Assumption 1.3, the causal effect on enrolment, $ACE_{\mathcal{U}_{s_0}}(E)$, in Equation (1.1) is

the proportion of OFA compliers in $\mathcal{U}_{s_0}$:

$$\begin{aligned} ACE_{\mathcal{U}_{s_0}}(E) &= Pr\left(B_i = OD_{*0} \mid i \in \mathcal{U}_{s_0}\right) + Pr\left(B_i = OD_{*1} \mid i \in \mathcal{U}_{s_0}\right) \\ &= \frac{1}{N_{\mathcal{U}_{s_0}}} \sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i = OC_{*0}\} + \frac{1}{N_{\mathcal{U}_{s_0}}} \sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i = OC_{*1}\}. \end{aligned}$$

1.4.2 Principal Causal Effects

Because principal strata are not affected by assignment, we can define causal effects conditional on the principal strata, known as principal causal effects.
Principal stratification with respect to the bivariate variable (E_i, C_i) makes it clear that a well-defined real value for the average causal effect of OFA on students' academic career exists only for the subgroup of always enrolled-students belonging to $\mathcal{U}_{s_0}$, because for the other types of students either $C_i(0) = *$ or $C_i(1) = *$. We refer to it as the survivor average causal effect (SACE) on students' academic career:

$$\begin{aligned} SACE_{\mathcal{U}_{s_0}}(C) &= \frac{\sum_{i \in \mathcal{U}_{s_0}} \left[C_i(1) - C_i(0)\right] \mathbb{I}\{B_i \in \{AE_{11}, AE_{00}, AE_{01}, AE_{10}\}\}}{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i \in \{AE_{11}, AE_{00}, AE_{01}, AE_{10}\}\}} \\ &= \frac{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i \in AE_{01}\} - \sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i \in AE_{10}\}}{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i \in \{AE_{11}, AE_{00}, AE_{01}, AE_{10}\}\}} \end{aligned} \tag{1.2}$$

where the second equality follows because for $i \in \mathcal{U}_{s_0}$, $C_i(1) - C_i(0) = 0$ if $B_i = AE_{11}, AE_{00}$; $C_i(1) - C_i(0) = 1 - 0$ if $B_i = AE_{01}$; and $C_i(1) - C_i(0) = 0 - 1$ if $B_i = AE_{10}$.
Similarly, a well-defined real value for the average causal effect of OFA on University credits, $Y_{i,UC}$, exists only for the subgroup of always-enrolled students belonging to $\mathcal{U}_{s_0}$. We refer to it as survivor average causal effect on University credits:

$$SACE_{\mathcal{U}_{s_0}}(UC) = \frac{\sum_{i \in \mathcal{U}_{s_0}} \left[Y_{i,UC}(1) - Y_{i,UC}(0)\right] \mathbb{I}\{B_i \in \{AE_{11}, AE_{00}, AE_{01}, AE_{10}\}\}}{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i \in \{AE_{11}, AE_{00}, AE_{01}, AE_{10}\}\}}. \tag{1.3}$$

The survivor average causal effect on University credits, $SACE_{\mathcal{U}_{s_0}}(UC)$, can be written as weighted average of survivor average causal effects on University credits across the four types of always-enrolled students (Never-active, OFA-inactive, OFA-active and Always-active students):

$$\begin{aligned} SACE_{\mathcal{U}_{s_0}}(UC) &= SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{00}) \cdot \pi_{AE_{00}|AE} + SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{01}) \cdot \pi_{AE_{01}|AE} + \\ &\quad SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{10}) \cdot \pi_{AE_{10}|AE} + SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11}) \cdot \pi_{AE_{11}|AE}, \end{aligned}$$

where

$$\pi_{AE_{\underline{c}}|AE} = \frac{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i = AE_{\underline{c}}\}}{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i \in \{AE_{11}, AE_{00}, AE_{01}, AE_{10}\}\}}$$

is the proportion of students of type $AE_{\underline{c}}$ among always-enrolled students, for $\underline{c} \in \{0,1\}^2$.

It is worth noting that:

$$\begin{aligned}
SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{00}) &= 0 \\
SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{10}) &= 0 - \mathbb{E}[Y_{i,UC}(0) \mid i \in \mathcal{U}_{s_0}] \\
SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{01}) &= \mathbb{E}[Y_{i,UC}(1) \mid i \in \mathcal{U}_{s_0}] - 0 \\
SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11}) &= \mathbb{E}[Y_{i,UC}(1) \mid i \in \mathcal{U}_{s_0}] - \mathbb{E}[Y_{i,UC}(0) \mid i \in \mathcal{U}_{s_0}]
\end{aligned}$$

Thus, we are particularly interested in $SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$, the average causal effect of OFA on UCs for always-active students for which both $Y_{i,UC}(0) > 0$ and $Y_{i,UC}(1) > 0$. Finally we are interested in the causal effect of OFA on the average grade at the end of the first academic year. A well-defined real value for this causal effect exists only for always-active students because both $Y_{i,G}(1)$ and $Y_{i,G}(0)$ are defined on $\mathbb{R}$ only for this group. Formally, we focus on the survivor average causal effect of OFA on average grade for always-active students:

$$SACE_{\mathcal{U}_{s_0}}(G) = \frac{\sum_{i \in \mathcal{U}_{s_0}} [Y_{i,G}(1) - Y_{i,G}(0)] \, \mathbb{I}\{B_i = AE_{11}\}}{\sum_{i \in \mathcal{U}_{s_0}} \mathbb{I}\{B_i = AE_{11}\}}. \tag{1.4}$$

1.5 Local Randomization

In order to draw inference on the causal effects of interest we need to introduce some assumption on the treatment assignment mechanism. Following Li et al. (2015) we formalize the concept of a RD design as local randomized experiment invoking the following assumption:

Assumption 1.4. (Local Randomization). For each $i \in \mathcal{U}_{s_0}$,

$$Pr\left(S_i \mid E_i(0), E_i(1), C_i(0), C_i(1), \mathbf{Y}_i(0), \mathbf{Y}_i(1), \mathbf{X}_i\right) = Pr\left(S_i\right).$$

Assumption 1.4 implies that for each unit within the sub-population, $i \in \mathcal{U}_{s_0}$, a Bernoulli trial has been conducted, with individual assignment probabilities depending only on the distribution of the forcing variable, not on either the potential outcomes or pre-treatment variables: $Pr\left(S_i \geq s_0 \mid E_i(0), E_i(1), C_i(0), C_i(1), \mathbf{Y}_i(0), \mathbf{Y}_i(1), \mathbf{X}_i\right) = Pr\left(S_i \geq s_0\right) =$ $= Pr\left(Z_i = 1\right)$. In other words, Assumption 1.4 implies that the treatment is randomly assigned in some small neighbourhood, $\mathcal{U}_{s_0}$, of s_0.

In our study the local randomization assumption appears to be reasonable. It is plausible to believe that for students with a value of the test score falling in a small neighbourhood of the threshold, the probability of falling on both sides of the threshold depends neither on the potential outcomes nor on pre-treatment variable. Moreover we can reasonably consider extremely difficult, if not impossible, for students to manipulate the value of the test score in order to end up on the right side of the threshold. Indeed participants have no information on the questions included in the test until the day of the test, the test is anonymous and related to participant only through a bar code, and the correction is

carried out by computers. Local RD-SUTVA (Assumption 1.2) and local randomization (Assumption 1.4) are reasonable assumptions for units belonging to $\mathcal{U}_{s_0}$, but may be debatable for subjects with a value of the test score very far-away from the threshold. For subjects whose score is either very low or very high, (1) RD-SUTVA may fail to hold because potential outcomes may directly depend on the value of the test score; and (2) local randomization may fail to hold because the test score may depend on some unobserved variable, such as motivation and ability, which may be related to the subsequent decision of enrolling and to the subsequent academic career and performances irrespective of the value of the test score.

1.6 Inference on Local Principal Causal Effects for a Sub-population $\mathcal{U}_{s_0}$

1.6.1 Selection of a Sub-population $\mathcal{U}_{s_0}$

Assumptions 1.1, 1.2 and 1.4 imply that there exists at least one sub-population $\mathcal{U}_{s_0}$ where the treatment (being exempt from OFA versus not being exempt from OFA) has been assigned according to a classical randomized experiment. Therefore if at least a true sub-population $\mathcal{U}_{s_0}$ were known, we could draw inference on causal effects for the sub-population $\mathcal{U}_{s_0}$ using standard methods for analysing randomized experiments, possibly suffering from post-treatment complications such as truncation by death (e.g. Zhang and Rubin, 2003; Rubin, 2006; Imbens and Rubin, 2015; Mattei and Mealli, 2007). Unfortunately, in practice, the true sub-populations $\mathcal{U}_{s_0}$ are usually unknown, and we need to select at least one sub-population $\mathcal{U}_{s_0}$.
In principle, a sub-population may come in any shape or form. Following Li et al. (2015), we focus on symmetric intervals about s_0, invoking the following assumption:

Assumption 1.5. (Shape of the Overlap Set). There exists $h > 0$ such that for each $\epsilon > 0$, $Pr(s_0 - h \leq S_i \leq s_0 + h) > 1 - \epsilon$ for each $i \in \mathcal{U}_{s_0}$.

Under Assumption 1.5, selecting suitable sub-populations $\mathcal{U}_{s_0}$ for causal inference means selecting values for the bandwidth, h, such that Assumptions 1.1-1.4 hold for the sub-group of units with values of the test score S_i in $[s_0 - h, s_0 + h]$. We adopt the approach proposed by Li et al. (2015), which relies on the local randomization assumption (Assumption 1.4).
Assumption 1.4 implies that within a sub-population $\mathcal{U}_{s_0}$ all observed and unobserved covariates are well balanced between the two sub-samples defined by the OFA status, that is, between participants who are assigned to OFA and participants who are not assigned to OFA have similar characteristics. Therefore, any test of the null hypothesis of no effect of assignment on covariates for units in $\mathcal{U}_{s_0}$ should fail to reject the null. A rejection can be interpreted as evidence against the local randomization assumption. If we are not able to reject the null, we can consider $\mathcal{U}_{s_0}$ as a suitable sup-population for causal inference under the assumption that all relevant variables known (or believed) to be related to both treatment assignment and the potential outcomes are observed.

Assessing balance in the observed covariates raises problems of multiple comparisons. We account for multiplicities using a Bayesian hierarchical mixed model, where we use a mixture for the prior distribution of the eligibility parameters by assigning a point mass on equality of the means of the covariates between subjects who are exempt from OFA and subjects who are not (Berry and Berry, 2004; Li et al., 2015; Mattei and Mealli, 2016). See Appendix 1.9 for technical details on model assumptions and specification of the prior distributions and for computational details.

This Bayesian procedure provides a measure of the risk (posterior probability) that a chosen interval around s_0, defines a sub-population of units that does not exactly matches any true $\mathcal{U}_{s_0}$, including units for which Assumptions 1.1-1.4 do not hold.

We apply this Bayesian procedure for assessing the balance of the covariates between the two groups defined by the OFA status for various sub-populations defined by different bandwidths, h. Table 1.1 shows the posterior probabilities that the covariates have the same distributions between the group of test participants who are exempt from OFA and the group of test participants who are not exempt from OFA, for bandwidths $h = 2.0, 2.5, 3.0, 3.5, 4.0$. As we can these probabilities are greater than 30% for $h = 2.0$ and $h = 2.5$. For larger bandwidths there is some evidence that the two groups of test participants defined by the OFA status differs in some background characteristics, making the local randomization assumption untenable. Specifically we find that the categorical variable "high school type" is not well balanced between the two groups of test participants, with a very low posterior probability for the category *Science* for values of h grater than 2.5. These results lead us to focus on the sub-population of test participants with a value of the test score falling in a neighbourhood of $h = 2.5$ around the threshold $s_0 = 16$: $\mathcal{U}_{s_0} = \{i \in \mathcal{U} : 13.5 \leq S_i \leq 18.5\}$. Because the score ranges between -20 and 80, scores in the interval [13.5; 18.5] correspond to percentages of success ranging between 13.5% and 18.5%.

Unfortunately there exists a non zero probability that the chosen interval around the threshold defines a sub-population that includes units who do not belong to any true sub-population, $\mathcal{U}_{s_0}$. In order to account for the presence of these units, the analyses are conducted conditioning on both the value of the forcing variable (the test score) and covariates.

1.6.2 Parametric Models

For the units within the selected sub-population $\mathcal{U}_{s_0}$, we conduct causal inference using a model-based Bayesian principal stratification analysis under Assumptions 1.1-1.4. The structure for Bayesian principal stratification inference was first developed in Imbens and Rubin (1997) for the special case of non-compliance. Building on Imbens and Rubin (1997) and Li et al. (2015), we assume parametric models for the conditional distribution of the principal stratum membership defined by $B_i = (\underline{E}_i, \underline{C}_i)$ given the test score and pre-treatment variables, and for the conditional distribution of potential outcomes for students' academic performances, $\mathbf{Y}_i(0)$ and $\mathbf{Y}_i(1)$, given the test score, pre-treatment

variables and principal stratum membership. All distributions are parametrized so that conditional on a general parameter, denoted by $\boldsymbol{\theta}$, the model has an independent and identical distribution (i.i.d.) structure.

Formally, we use the following linear transformation of the forcing variable: $\tilde{S}_i = (S_i - s_0)/10$. We factorize the conditional distribution of $B_i = (\underline{E}_i, \underline{C}_i)$ given the test score and the covariates as

$$Pr\left(B_i = b \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}\right) = Pr\left(\underline{E}_i = \underline{e} \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}\right) \cdot Pr\left(\underline{C}_i = \underline{c} \mid \underline{E}_i = \underline{e}, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}\right)$$

for $b \in \{NE_{**}, AE_{00}, AE_{11}, AE_{01}, AE_{10}, OC_{*0}, OC_{*1}\}$; $\underline{e} \in \{NE, AE, OC\}$; and $\underline{c} \in \{(*,*),$ $(*,0)$, $(*,1)$, $(0,0)$, $(1,1)$, $(0,1)$, $(1,0)\}$.

We specify a conditional probit model for the principal strata defined by the enrolment status,

$\underline{E}_i \in \{AE, NE, OC\}$, given $\tilde{S}_i = (S_i - s_0)/10$ and $\mathbf{X}_i$

$$\begin{array}{rcll}
\pi_i(AE) & \equiv & Pr(\underline{E}_i = AE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) & = & Pr(E_i^*(AE) \leq 0 \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\
\pi_i(NE) & \equiv & Pr(\underline{E}_i = NE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) & = & Pr(E_i^*(AE) > 0,\, E_i^*(NE) \leq 0 \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\
\pi_i(OC) & \equiv & Pr(\underline{E}_i = OC \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) & = & 1 - Pr(E_i = NE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) - \\
 & & & & Pr(E_i = AE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}),
\end{array}$$

where $\epsilon_{i,AE} \sim \mathcal{N}(0,1)$ and $\epsilon_{i,NE} \sim \mathcal{N}(0,1)$, and $\epsilon_{i,AE}$ and $\epsilon_{i,NE}$ are independent.

Note that $Pr(\underline{C}_i = (*,*) \mid \underline{E}_i = NE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) = 1$ and $Pr(\underline{C}_i = \underline{c} \mid \underline{E}_i = NE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) = 0$ for $\underline{c} \equiv (c_0, c_1) \neq (*,*)$.

We specify four conditional probit models for $\underline{C}_i$, given $\underline{E}_i = AE$, $\tilde{S}_i$ and $\mathbf{X}_i$:

$$\begin{array}{rcl}
\pi_i(00 \mid AE) & \equiv & Pr(\underline{C}_i = (0,0) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\
 & = & Pr(C_i^*(0,0) \leq 0 \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\
\pi_i(01 \mid AE) & \equiv & Pr(\underline{C}_i = (0,1) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\
 & = & Pr(C_i^*(0,0) > 0,\, C_i^*(0,1) \leq 0 \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\
\pi_i(11 \mid AE) & \equiv & Pr(\underline{C}_i = (1,1) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\
 & = & Pr(C_i^*(0,0) > 0,\, C_i^*(0,1) > 0, C_i^*(1,1) \leq 0 \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\
\pi_i(10 \mid AE) & \equiv & Pr(\underline{C}_i = (1,0) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\
 & = & 1 - \pi_i(00 \mid AE) - \pi_i(01 \mid AE) - \pi_i(11 \mid AE)
\end{array}$$

where $\epsilon_{i,AE_{\underline{c}}} \sim \mathcal{N}(0,1)$, independently, for $\underline{c} = (0,0), (0,1), (1,1)$.

We have $Pr(\underline{C}_i = (*,*) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) = Pr(\underline{C}_i = (*,0) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) =$ $Pr(\underline{C}_i = (*,1) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) = 0$.

Similarly we have $Pr(\underline{C}_i = \underline{c} \mid \underline{E}_i = OC, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) = 0$ for $\underline{c} \in \{(*,*),\ (0,0),\ (1,1),$ $(0,1),\ (1,0)\}$; $Pr(\underline{C}_i = \underline{c} \mid \underline{E}_i = OC, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) > 0$ only for $\underline{c} \in \{(*,0), (*,1)\}$ under Assumption 1.3. Therefore we specify a probit model for $\underline{C}_i$, given $\underline{E}_i = OC$, $\tilde{S}_i$ and $\mathbf{X}_i$:

$$\underline{C}_i = (*,0) \mid \underline{E}_i = OC \quad if \quad C_i^*(*,0) = \beta_{OC_{*0}}^{(0)} + \tilde{S}_i \beta_{OC_{*0}}^{(S)} + \mathbf{X}_i' \boldsymbol{\beta}_{OC_{*0}}^{(X)} + \epsilon_{i,OC_{*0}} \geq 0$$

where $\epsilon_{i,OC_{*0}} \sim \mathcal{N}(0,1)$.

We now need to specify a parametric model for the conditional distribution of potential outcomes for students' academic performances, $\mathbf{Y}_i(z) = (Y_{i,UC}(z), Y_{i,G}(z))$, $z = 0, 1$, given the forcing variable, $\tilde{S}_i$, pre-treatment variables, $\mathbf{X}_i$, and principal stratum membership, B_i. To this end, it is useful to factorize the conditional distribution of $\mathbf{Y}_i(z) = (Y_{i,UC}, Y_{i,G})$ as product of the marginal distribution of University credits, $Y_{i,UC}(z) \mid B_i, \tilde{S}_i, \mathbf{X}_i$, and the conditional distribution of average grade given University credits, $Y_{i,G}(z) \mid Y_{i,UC}(z), B_i, \tilde{S}_i, \mathbf{X}_i$. We assume that conditional on $\tilde{S}_i$, $\mathbf{X}_i$, B_i and $\boldsymbol{\theta}$, the potential outcomes $\mathbf{Y}_i(0)$ and $\mathbf{Y}_i(1)$ are independent[4].

University credits is a discrete variable with support in $\{*\} \cup \{0, 6, 9, 12, \ldots, 60\}$. We have that $Y_{i,UC}(z)$ is not defined for test participants who would not enrol under OFA status z, $z = 0, 1$, and $Y_{i,UC}(z) = 0$ for test participants who would enrol but would be inactive under OFA status z, $z = 0, 1$. Thus, $Pr\left(Y_{i,UC}(z) = * \mid B_i = b, \tilde{S}_i, \mathbf{X}_i\right) = 1$ for $(z,b) = (0, NE_{**}), (1, NE_{**}), (0, OC_{*0}), (0, OC_{*1})$; and $Pr\left(Y_{i,UC}(z) = 0 \mid B_i = b, \tilde{S}_i, \mathbf{X}_i\right) = 1$ for $(z,b) = (0, AE_{00}), (1, AE_{00}), (0, AE_{01}), (1, AE_{10}), (1, OC_{*0})$. For $(z,b) \in \{(0, AE_{11}), (1, AE_{11}), (1, AE_{01}), (0, AE_{10}), (1, OC_{*1})\}$, the variable University credits takes on values in $\{6, 9, 12, \ldots, 60\}$. We specify a Poisson regression model with the canonical link function for the following linear transformation of University credits: $\widetilde{Y}_{i,UC}(z) = \dfrac{Y_{i,UC}(z)}{3} - 2$, which has support $\{0, 1, 2, \ldots\}$.

Formally, we assume that $\widetilde{Y}_{UC_i}(z) \mid B_i = b, S_i, \mathbf{X}_i; \boldsymbol{\theta} \sim Poisson(\lambda_i(z,b))$, where

$$\log(\lambda_i(z,b)) = \gamma_{z,b}^{(0)} + \tilde{S}_i \gamma_{z,b}^{(S)} + \mathbf{X}_i' \boldsymbol{\gamma}_{z,b}^{(X)},$$

for $(z,b) \in \{(0, AE_{11}), (0, AE_{01}), (1, AE_{11}), (1, AE_{10}), (1, OC_{*0})\}$.

Potential outcomes for average grade, $Y_{i,G}(z)$, $z = 0, 1$, have support on $\{*\} \cup [18, 30]$, and we have $Pr\left(Y_{i,G}(z) = * \mid Y_{i,UC}(z), B_i = b, \tilde{S}_i, \mathbf{X}_i\right) = 1$ for $(z,b) \in \{(0, NE_{**}), (1, NE_{**}), (0, OC_{*0}), (1, OC_{*0}), (0, OC_{*1}), (0, AE_{00}), (1, AE_{00}), (0, AE_{01}), (1, AE_{10})\}$.

For $(z,b) \in \{(0, AE_{11}), (1, AE_{11}), (1, AE_{01}), (0, AE_{10}), (1, OC_{*1})\}$, $Y_{i,G}(z)$ is a well defined random variable with support on $[18, 30]$.

We model these five outcome distributions using truncated log-normal regressions: let $\widetilde{Y}_{i,G}(z) = \log(Y_{i,G}(z))$, then $\widetilde{Y}_{i,G}(z) \mid Y_{i,UC}, B_i = b, \tilde{S}_i, \mathbf{X}_i \sim \mathcal{TN}(\tau_i(z,b), \sigma_{z,b}^2)$, with support $[\log(18 - 1/3), \log(30 + 1/3)]$, where

$$\tau_i(z,b) = \eta_{z,b}^{(0)} + \tilde{S}_i \eta_{z,b}^{(S)} + \widetilde{Y}_{i,UC}(z) \eta_{z,b}^{(UC)} + \mathbf{X}_i' \boldsymbol{\eta}_{z,b}^{(X)}.$$

Because the sample size is relatively small, in order to avoid imprecise estimates, we assume equality of some slope coefficients in the outcomes' models: $\boldsymbol{\gamma}^{(X)} \equiv \boldsymbol{\gamma}_{z,b}^{(X)}$ for $(z,b) \in \{(0, AE_{11}), (1, AE_{11}), (1, AE_{01}), (0, AE_{10}), (1, OC_{*0}), (1, OC_{*1})\}$; and $\boldsymbol{\eta}^{(X)} \equiv \boldsymbol{\eta}_{z,b}^{(X)}$ for $(z,b) \in \{(0, AE_{11}), (1, AE_{11}), (1, AE_{01}), (0, AE_{10}), (1, OC_{*1})\}$. We also impose the

[4]Because $\mathbf{Y}_i(0)$ and $\mathbf{Y}_i(1)$ are never jointly observed, we cannot expect to learn anything about the partial correlation between them from the data given $\tilde{S}_i$, $\mathbf{X}i$, B_i and $\boldsymbol{\theta}$ (Rubin, 1978; Imbens and Rubin, 1997).

following prior equalities: $\sigma^2_{AE} \equiv \sigma^2_{0,AE_{11}} = \sigma^2_{1,AE_{11}} = \sigma^2_{0,AE_{10}} = \sigma^2_{1,AE_{01}}$ in the truncated log-normal regression models for average grade for always-enrolled students..

We assume that parameters are a *priori* independent, and we use multivariate normal prior distributions for the regression coefficients and Scaled-Inverse-χ^2 distributions for the parameters σ^2_{AE} and $\sigma^2_{1,OC}$ of the truncated log-normal models. See Appendix 1.9 for details on the specification of the prior distributions. In Appendix 1.9 we also assess the sensitivity of the results with respect to the specification of the prior distributions for the parameters σ^2_{AE} and $\sigma^2_{1,OC}$ of the log-normal models, by using uninformative priors. The results appear to be robust changing only slightly under uninformative priors for σ^2_{AE} and $\sigma^2_{1,OC}$.

It is worth noting that although we specify probit and conditional probit models for binary and categorical covariates, respectively, alternative specification, such as logit and multinomial logit models, could be used, too (e.g., Hirano et al., 2000). We opt for probit and conditional probit models for computational convenience given that, conditional on latent variables, the full conditional distributions of the parameters of these models have a closed form, from which we can easily sample. Nevertheless, we expect that results would be very similar.

The posterior distribution has a complex form involving mixtures due to the fact that the observed groups defined by the observed OFA status, Z, the observed status, E^{obs}, and the observed academic career status, C^{obs}, comprise mixtures of principal strata (see Table 1.3). Indeed we cannot generally observe the principal stratum which a unit belongs to, because for each unit only one potential outcome is observed for each post-treatment variable, depending on the treatment actually received. We deal with these computational issues developing an MCMC algorithm which uses the Data Augmentation (DA) method of Tanner and Wong (1987) to impute the missing $E^{mis}_i = E_i(1 - Z_i)$ and $C^{mis}_i = C_i(1 - Z_i)$. Results shown in the next section are based on a chain which was run for 25 000 iterations, burning the first 5 000 iterations.

1.7 Causal Inference for the Sub-population $\mathcal{U}_{s_0}$

Tables 1.4 and 1.5 show the posterior medians and 95% posterior credible intervals (PCIs) of the posterior distributions of the principal strata proportions and of the causal estimands under monotonicity (Assumption 1.3). The estimated proportions of the principal strata suggest that most of test participants who obtain a test score between 13.5 and 18.5 are either always-enrolled students or never-enrolled students: there are around 47% always-enrolled students, more than 40% never-enrolled students and around 11% OFA compliers. Under monotonicity the proportion of OFA compliers is the causal effect of the OFA status on the decision to enrol, $ACE_{\mathcal{U}_{s_0}}(E)$, which is not negligible (see also Table 1.5). Among OFA compliers 71% are active students, that is, students who would take and pass at least one exam if they did not received OFA. Among always-enrolled students, 43% are always-active students, 26% are OFA-inactive students, around 19% are OFA-active students and 12% are never-active students.

The posterior median of the survivor causal effect on academic career for always-enrolled students, $SACE_{\mathcal{U}_{s_0}}(C)$, is equal to 3% and the 95% posterior credible interval ranges from -0.003 to 0.071. The posterior probability that $SACE_{\mathcal{U}_{s_0}}(C)$ is positive is approximately 96%. A positive value of $SACE_{\mathcal{U}_{s_0}}(C)$ suggests that being exempt from OFA increases the probability of being an active student. Thus, there appears to be some evidence that being exempt from OFA helps students to start their academic career taking and pass at least one exam.

The posterior median of the survivor causal effect on University Credits for always-enrolled students, $SACE_{\mathcal{U}_{s_0}}(UC)$, is 1.634 suggesting that always-enrolled students who are exempt from OFA award on average about one more credit than always-enrolled students who are not exempt from OFA. Nevertheless the 95% posterior credible interval of $SACE_{\mathcal{U}_{s_0}}(UC)$, covers zero and is rather wide (95% PCI $= [-7.378, 10.614]$), so that $SACE_{\mathcal{U}_{s_0}}(UC)$ is no significant at the 5% level. The posterior median of the survivor causal effect on University Credits for the sub-group of always-enrolled students who are always active irrespective of their OFA status, $SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$, is 3.137, and thus, is positive and larger than the posterior median of the survivor causal effect on University Credits for all the always-enrolled students. This effect, however, is estimated very imprecisely: the 95% posterior credible interval, $[-14.848, 22.414]$, is very wide and spread around zero. It is worth noting that freshmen of the school of Engineering are expected to award 60 UCs by the end of the first year, taking exams to which are generally assigned 6, 9 or 12 UCs. Therefore our findings suggest that being exempt from OFA versus not being exempt from OFA does not significantly affect the number of passed exams: students take approximately the same number of exams irrespective of their OFA status.

The estimated survivor average causal effect on average grade for always-active students, $SACE_{\mathcal{U}_{s_0}}(G)$, is positive, but small and statistically negligible: the posterior median is an increase of the average grade of just 0.41 points, and the 95% posterior credible interval covers zero. As we can seen in the right graph in Figure 1.1 the posterior distribution of $SACE_{\mathcal{U}_{s_0}}(G)$ is almost symmetric around zero. Figure 1.2 shows the joint distribution of $SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$ and $SACE_{\mathcal{U}_{s_0}}(G)$, which suggests that there exists a very weak (if any) correlation between the effects of the OFA status on UCs and on average grade for always-active students.

1.8 Conclusions

We focus on assessing casual effects of the OFA status on academic career and performances at the end of the first academic year for students who take the entrance test for the School of Engineering of a given Italian State University. The OFA assignment rule appeals to a (sharp) RD design with the entrance exam score acting as forcing variable. Following Li et al. (2015), we formally describe our RD design as a local randomized experiment. For the sub-population of units for which the critical assumptions hold, we define the causal estimands of interest using the framework of

principal stratification, which allow us to properly deal with the problem of truncation by "death" that affects our outcome variables: academic career status and University credits, which are truncated by not enrolment, and average grade, which is truncated by enrolment and by inactivity. We conduct causal inference on local principal causal effects using a model-based Bayesian approach. In our application we find little evidence that being exempt from OFA has any beneficial effects on students' academic career and performances for the sub-population of test participants who obtain a test score between 13.5 and 18.5. The strongest evidence is that being exempt from OFA appears to increase the probability of being an active students for always-enrolled students. Being exempt from OFA appears also to increase the number of UCs for always-enrolled students, although the estimated effect is small and with large posterior variability. We interpret this result as evidence that students who are assigned OFA because they obtain a test score less than 16 points may delay the start of their academic career and thus, have a higher risk to be inactive or award a lower number of UCs during the first academic year than students who are not assigned OFA, because they must comply with additional compulsory learning, by attending an extra course and taking an extra exam. Indeed, we find a slightly larger effect of OFA on number of UCs for always-active students, even if this effect is estimated very imprecisely. Therefore, our analysis suggests that for the sub-population of test participants who obtain a test score falling in a symmetric interval around the threshold, $s_0 = 16$, defined by a bandwidth equal to $h = 2.5$, being exempt from OFA may encourage always-enrolled students to promptly activate their academic career and to award a higher number of UCs, affecting students' academic performances from a quantitative standpoint. Nevertheless, being exempt from OFA does not seem significantly affect the quality of always-active students' academic performances as measured by average grade.
In our application study we focus on a school of Engineering of a specific University. It would be interesting to consider all schools of Engineering that adopted the same entrance exam in 2011. However, this extension raises some challenges because different schools use different thresholds to define the OFA status. A valuable topic for future research is to investigate the heterogeneity of the effect with respect to the threshold.

1.9 Appendix

1.9.1 Bayesian Selection of the Subpopulations

We used a hierarchical Bayesian model for assessing the balance of the covariates between the two groups defined by the OFA status: the exempt from OFA group and the not exempt from OFA group. The posterior distributions of the parameters are obtained from Markov chain Monte Carlo (MCMC) methods.
Let X_1, X_2, X_3, X_4, X_5, and X_6 denote the six observed covariates where X_1, X_2, X_3, and X_6 are binary variables for gender, high school year; irregular high school career; and living far away from the campus; X_4 is a categorical variable for high school type with 4 categories; and X_5 is high school grade, a continuous variable with support in

[60, 100]. Under the model specification described in Section 5.1 of the main text, the joint posterior distribution of the parameters is

$$
\begin{aligned}
&p(\sigma^2_{\gamma_0}, \mu_{\gamma_0}, \sigma^2_{\gamma_1}, \mu_{\gamma_1}, \pi_{\gamma_1}, \sigma^2_5, \gamma_{01}, \gamma_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \gamma_{06}, \gamma_{11}, \gamma_{12}, \boldsymbol{\gamma}_{14}, \gamma_{13}, \gamma_{15}, \gamma_{16} \mid \mathbf{X}, \mathbf{Z}) \propto \\
&\quad p(\sigma^2_{\gamma_0}, \mu_{\gamma_0}, \sigma^2_{\gamma_1}, \mu_{\gamma_1}, \pi_{\gamma_1}) \times \\
&\quad p(\sigma^2_5, \gamma_{01}, \gamma_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \gamma_{06}, \gamma_{11}, \gamma_{12}, \boldsymbol{\gamma}_{14}, \gamma_{13}, \gamma_{15}, \gamma_{16} \mid \sigma^2_{\gamma_0}, \mu_{\gamma_0}, \sigma^2_{\gamma_1}, \mu_{\gamma_1}, \pi_{\gamma_1}) \times \\
&\quad \mathcal{L}(\sigma^2_5, \gamma_{01}, \gamma_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \gamma_{06}, \gamma_{11}, \boldsymbol{\gamma}_{12}, \gamma_{14}, \gamma_{13}, \gamma_{15}, \gamma_{16}; \mathbf{X}, \mathbf{Z})
\end{aligned}
$$

where

$$
\begin{aligned}
&p(\sigma^2_{\gamma_0}, \mu_{\gamma_0}, \sigma^2_{\gamma_1}, \mu_{\gamma_1}, \pi_{\gamma_1}) = \\
&\quad \frac{\underline{b}_{\gamma_0}^{\underline{a}_{\gamma_0}}}{\Gamma(\underline{a}_{\gamma_0})} (\sigma^2_{\gamma_0})^{-(\underline{a}_{\gamma_0}-1)} \exp\left\{-\frac{\underline{b}_{\gamma_0}}{\sigma^2_{\gamma_0}}\right\} \times \frac{1}{\sqrt{2\pi\underline{\sigma}^2_{\gamma_0}}} \exp\left\{-\frac{1}{2\underline{\sigma}^2_{\gamma_0}} \left(\mu_{\gamma_0} - \underline{\mu}_{\gamma_0}\right)^2\right\} \times \\
&\quad \frac{\underline{b}_{\gamma_1}^{\underline{a}_{\gamma_1}}}{\Gamma(\underline{a}_{\gamma_1})} (\sigma^2_{\gamma_1})^{-(\underline{a}_{\gamma_1}-1)} \exp\left\{-\frac{\underline{b}_{\gamma_1}}{\sigma^2_{\gamma_1}}\right\} \times \frac{1}{\sqrt{2\pi\underline{\sigma}^2_{\gamma_1}}} \exp\left\{-\frac{1}{2\underline{\sigma}^2_{\gamma_1}} \left(\mu_{\gamma_1} - \underline{\mu}_{\gamma_1}\right)^2\right\} \times \\
&\quad \frac{\Gamma(\underline{a}_\pi + \underline{b}_\pi)}{\Gamma(\underline{a}_\pi)\Gamma(\underline{b}_\pi)} \pi_{\gamma_1}^{\underline{a}_\pi - 1} (1 - \pi_{\gamma_1})^{\underline{b}_\pi - 1}
\end{aligned}
$$

$$
\begin{aligned}
&p(\sigma^2_5, \gamma_{01}, \boldsymbol{\gamma}_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \boldsymbol{\gamma}_{06}, \gamma_{11}, \boldsymbol{\gamma}_{12}, \boldsymbol{\gamma}_{14}, \gamma_{13}, \gamma_{15}, \boldsymbol{\gamma}_{16} \mid \sigma^2_{\gamma_0}, \mu_{\gamma_0}, \sigma^2_{\gamma_1}, \mu_{\gamma_1}, \pi_{\gamma_1}) = \\
&\quad \frac{b^a}{\Gamma(a)} (\sigma^2_5)^{-(a-1)} \exp\left\{-\frac{b}{\sigma^2_5}\right\} \times \prod_{j=1,2,3,5,6} \frac{1}{\sqrt{2\pi\sigma^2_{\gamma_0}}} \exp\left\{-\frac{1}{2\sigma^2_{\gamma_0}} \left(\gamma_{0j} - \mu_{\gamma_0}\right)^2\right\} \times \\
&\quad \prod_{k=1}^{3} \frac{1}{\sqrt{2\pi\sigma^2_{\gamma_0}}} \exp\left\{-\frac{1}{2\sigma^2_{\gamma_0}} \left(\gamma_{04}^{(k)} - \mu_{\gamma_0}\right)^2\right\} \times \\
&\quad \prod_{j=1,2,3,5,6} \left[\pi_{\gamma_1} \mathbb{I}(\{\gamma_{1j} = 1\}) + (1 - \pi_{\gamma_1}) \frac{1}{\sqrt{2\pi\sigma^2_{\gamma_1}}} \exp\left\{-\frac{1}{2\sigma^2_{\gamma_1}} \left(\gamma_{1j} - \mu_{\gamma_1}\right)^2\right\}\right] \times \\
&\quad \prod_{k=1}^{3} \left[\pi_{\gamma_1} \mathbb{I}(\{\gamma_{14}^{(k)} = 1\}) + (1 - \pi_{\gamma_1}) \frac{1}{\sqrt{2\pi\sigma^2_{\gamma_1}}} \exp\left\{-\frac{1}{2\sigma^2_{\gamma_1}} \left(\gamma_{14}^{(k)} - \mu_{\gamma_1}\right)^2\right\}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathcal{L}(\sigma^2_5, \gamma_{01}, \gamma_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \gamma_{06}, \gamma_{11}, \gamma_{12}, \boldsymbol{\gamma}_{14}, \gamma_{13}, \gamma_{15}, \gamma_{16}; \mathbf{X}, \mathbf{Z}) = \\
&\quad \prod_i \prod_{j=1,2,3,6} \Phi\left((-1)^{\mathbb{I}\{X_{ij}=0\}} (\gamma_{0j} + \gamma_{1j} Z_i)\right) \times \frac{1}{\sqrt{2\pi\sigma^2_5}} \exp\left\{\frac{1}{\sigma^2_5} \left(X_{i5} - \gamma_{05} - \gamma_{15} Z_i\right)^2\right\} \times \\
&\quad \left[1 - \Phi\left(\gamma_{04}^{(1)} + \gamma_{14}^{(1)} Z_i\right)\right]^{\mathbb{I}\{X_{i4}=1\}} \left[\Phi\left(\gamma_{04}^{(1)} + \gamma_{14}^{(1)} Z_i\right) \left(1 - \Phi\left(\gamma_{04}^{(2)} + \gamma_{12}^{(2)} Z_i\right)\right)\right]^{\mathbb{I}\{X_{i4}=2\}} \\
&\quad \left[\Phi\left(\gamma_{04}^{(1)} + \gamma_{14}^{(1)} Z_i\right) \Phi\left(\gamma_{02}^{(2)} + \gamma_{14}^{(2)} Z_i\right) \left(1 - \Phi\left(\gamma_{04}^{(3)} + \gamma_{14}^{(3)} Z_i\right)\right)\right]^{\mathbb{I}\{X_{i4}=3\}} \\
&\quad \left[\Phi\left(\gamma_{04}^{(1)} + \gamma_{14}^{(1)} Z_i\right) \Phi\left(\gamma_{04}^{(2)} + \gamma_{14}^{(2)} Z_i\right) \Phi\left(\gamma_{04}^{(3)} + \gamma_{14}^{(3)} Z_i\right)\right]^{\mathbb{I}\{X_{i4}=4\}}
\end{aligned}
$$

The full conditional distributions for the hyper-parameters are

$$\begin{aligned}\sigma^2_{\gamma_0} &\mid \mu_{\gamma_0}, \gamma_{01}, \gamma_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \gamma_{06} \sim IG\left(\overline{a}_{\gamma_0}, \overline{b}_{\gamma_0}\right)\\ \mu_{\gamma_0} &\mid \sigma^2_{\gamma_0}, \gamma_{01}, \gamma_{02}, \gamma_{03}, \boldsymbol{\gamma}_{04}, \gamma_{05}, \gamma_{06} \sim N(\overline{\mu}_{\gamma_0}, \overline{\sigma}^2_{\gamma_0})\end{aligned}$$

where

$$\begin{aligned}\overline{a}_{\gamma_0} &= \underline{a}_{\gamma_0} + \frac{N_{\gamma_0}}{2}\\ \overline{b}_{\gamma_0} &= \underline{b}_{\gamma_0} + \frac{1}{2}\left[\sum_{j=1,2,3,5,6}(\gamma_{0j} - \mu_{\gamma_0})^2 + \sum_{k=1}^{3}(\gamma_{04}^{(k)} - \mu_{\gamma_0})^2\right]\\ \overline{\sigma}^2_{\gamma_0} &= \left(\frac{1}{\underline{\sigma}^2_{\gamma_0}} + \frac{N_{\gamma_0}}{\sigma^2_{\gamma_0}}\right)^{-1}\\ \overline{\mu}_{\gamma_0} &= \overline{\sigma}^2_{\gamma_0}\left[\frac{\underline{\mu}_{\gamma_0}}{\underline{\sigma}^2_{\gamma_0}} + \frac{1}{\sigma^2_{\gamma_0}}\left(\sum_{j=1,2,3,5,6}\gamma_{0j} + \sum_{k=1}^{3}\gamma_{04}^{(k)}\right)\right];\end{aligned}$$

$$\begin{aligned}\sigma^2_{\gamma_1} &\mid \mu_{\gamma_1}, \gamma_{11}, \gamma_{12}, \gamma_{13}, \boldsymbol{\gamma}_{14}, \gamma_{15}, \gamma_{16} \sim IG\left(\overline{a}_{\gamma_1}, \overline{b}_{\gamma_1}\right)\\ \mu_{\gamma_1} &\mid \sigma^2_{\gamma_1}, \gamma_{11}, \gamma_{12}, \gamma_{13}, \boldsymbol{\gamma}_{14}, \gamma_{15}, \gamma_{16} \sim N(\overline{\mu}_{\gamma_1}, \overline{\sigma}^2_{\gamma_1}),\end{aligned}$$

where

$$\begin{aligned}\overline{a}_{\gamma_1} &= \underline{a}_{\gamma_1} + \frac{N_{\gamma_1}}{2}\\ \overline{b}_{\gamma_1} &= \underline{b}_{\gamma_1} + \frac{1}{2}\left[\sum_{j=1,2,3,5,6}(\gamma_{1j} - \mu_{\gamma_1})^2 + \sum_{k=1}^{3}(\gamma_{14}^{(k)} - \mu_{\gamma_1})^2\right]\\ \overline{\sigma}^2_{\gamma_1} &= \left(\frac{1}{\underline{\sigma}^2_{\gamma_1}} + \frac{N_{\gamma_1}}{\sigma^2_{\gamma_1}}\right)^{-1}\\ \overline{\mu}_{\gamma_1} &= \overline{\sigma}^2_{\gamma_1}\left[\frac{\underline{\mu}_{\gamma_1}}{\underline{\sigma}^2_{\gamma_1}} + \frac{1}{\sigma^2_{\gamma_1}}\left(\sum_{j=1,2,3,5,6}\gamma_{1j} + \sum_{k=1}^{3}\gamma_{14}^{(k)}\right)\right];\end{aligned}$$

and

$$\pi_{\gamma_1} \mid \gamma_{11}, \gamma_{12}, \gamma_{13}, \boldsymbol{\gamma}_{14}, \gamma_{15}, \gamma_{16} \sim Beta(\overline{a}_{\pi}, \overline{b}_{\pi}),$$

where

$$\begin{aligned}\overline{a}_{\pi} &= \underline{a}_{\pi} + \left(\sum_{j=1,2,3,5,6}\mathbb{I}\{\gamma_{1j} = 0\} + \sum_{k=1}^{3}\mathbb{I}\{\gamma_{14}^{(k)} = 0\}\right)\\ \overline{b}_{\pi} &= \underline{b}_{\pi} + N_{\gamma_1} - \left(\sum_{j=1,2,3,5,6}\mathbb{I}\{\gamma_{1j} = 0\} + \sum_{k=1}^{3}\mathbb{I}\{\gamma_{14}^{(k)} = 0\}\right)\end{aligned}$$

with $\mathcal{N}_{\gamma_0} = 8$ and $\mathcal{N}_{\gamma_1} = 8$ are the number of $\gamma_0's$ and $\gamma_1's$.
The full conditional distributions for the parameters are as follows:

Parameters of the distribution of the continuous variable, X_5 (High school grade):

$$\sigma_5^2 \mid \gamma_{05}, \gamma_{15}, \mathbf{X}, \mathbf{Z} \sim IG\left(\overline{a}, \overline{b}\right) \qquad \gamma_{05} \mid \sigma_5^2, \gamma_{15}, \sigma_{\gamma_0}^2, \mu_{\gamma_0}, \mathbf{X}, \mathbf{Z} \sim N\left(\overline{\mu}_{\gamma_{05}}, \overline{\sigma}_{\gamma_{05}}^2\right)$$

where

$$\overline{a} = \underline{a} + \frac{N}{2} \qquad \overline{b} = \underline{b} + \frac{1}{2}\sum_i \left(X_{i5} - \gamma_{05} - \gamma_{15}Z_i\right)^2$$
$$\overline{\sigma}_{\gamma_{05}}^2 = \left(\frac{1}{\sigma_{\gamma_0}^2} + \frac{N}{\sigma_5^2}\right)^{-1} \qquad \overline{\mu}_{\gamma_{05}} = \overline{\sigma}_{\gamma_{05}}^2 \left[\frac{\mu_{\gamma_0}}{\sigma_{\gamma_0}^2} + \frac{\sum_i (X_{i5} - \gamma_{15}Z_i)}{\sigma_5^2}\right]$$

and

$$p\left(\gamma_{15} \mid \sigma_5^2, \gamma_{05}, \pi_{\gamma_1}, \sigma_{\gamma_1}^2, \mu_{\gamma_1}, \mathbf{X}, \mathbf{Z}\right) \propto \left(\frac{1}{\sqrt{2\pi\sigma_5^2}}\right)^N \exp\left\{-\frac{1}{2\sigma_5^2}\sum_i (X_{i5} - \gamma_{05} - \gamma_{15}Z_i)^2\right\} \times$$
$$\left[\pi_{\gamma_1}\mathbb{I}(\{\gamma_{15} = 0\}) + (1 - \pi_{\gamma_1})\frac{1}{\sqrt{2\pi\sigma_{\gamma_1}^2}} \exp\left\{-\frac{1}{2\sigma_{\gamma_1}^2}\left(\gamma_{15} - \mu_{\gamma_1}\right)^2\right\}\right]$$

Parameters of the distribution of the binary variables, X_1 (Sex), X_2 (High school Year), X_3 (Irregular Career) and X_6 (Far-away resident):
Let $X_{ij}^* \sim N\left(\gamma_{0j} + \gamma_{1j}Z_i, 1\right)$, $j = 1, 2, 3, 6$ a latent variable such that $Pr(X_{ij} = 1) = Pr(X_{ij}^* > 0)$. The distribution of $X_{ij}^* \mid \gamma_{0j}, \gamma_{1j}, \mathbf{X}, \mathbf{Z}$ is $\mathcal{N}(\gamma_{0j} + \gamma_{1j}Z_i, 1)$ truncated to the left of zero if $X_{ij} = 1$ and truncated to the right of zero if $X_{ij} = 0$.
We have

$$\gamma_{0j} \mid \gamma_{1j}, \sigma_{\gamma_0}^2, \mu_{\gamma_0}, \mathbf{X}, \mathbf{Z}, \mathbf{X}_j^* \sim N\left(\overline{\mu}_{\gamma_{0j}}, \overline{\sigma}_{\gamma_{0j}}^2\right)$$

where

$$\overline{\sigma}_{\gamma_{0j}}^2 = \left(\frac{1}{\sigma_{\gamma_0}^2} + N\right)^{-1} \qquad \overline{\mu}_{\gamma_{0j}} = \overline{\sigma}_{\gamma_{0j}}^2 \left[\frac{\mu_{\gamma_0}}{\sigma_{\gamma_0}^2} + \sum_i (X_{ij}^* - \gamma_{1j}Z_i)\right]$$

and

$$p\left(\gamma_{1j} \mid \gamma_{0j}, \pi_{\gamma_1}, \sigma_{\gamma_1}^2, \mu_{\gamma_1}, \mathbf{X}, \mathbf{Z}\right) \propto \Phi\left((-1)^{\mathbb{I}\{X_{ij}=0\}}(\gamma_{0j} + \gamma_{1j}Z_i)\right) \times$$
$$\left[\pi_{\gamma_1}\mathbb{I}(\{\gamma_{1j} = 0\}) + (1 - \pi_{\gamma_1})\frac{1}{\sqrt{2\pi\sigma_{\gamma_1}^2}} \exp\left\{-\frac{1}{2\sigma_{\gamma_1}^2}\left(\gamma_{1j} - \mu_{\gamma_1}\right)^2\right\}\right]$$

Parameters of the distribution of the categorical variables, X_4 (High school type):
Let $K = 4$ be the number of values/levels for X_4 and let $X_{i4}^{*(k)} \sim N\left(\gamma_{04} + \gamma_{14}Z_i, 1\right)$,

$K = 1, \ldots, K-1$, be independent latent variables such that $Pr(X_{i4} = 1) = Pr\left(X_{i4}^{*(1)} \leq 0\right)$, and $Pr(X_{i4} = k) = Pr\left(\cap_{\ell=1}^{k-1}\{X_{i4}^{*(\ell)} > 0\} \cap X_{i4}^{*(k)} \leq 0\right)$ for $k = 2, \ldots, K-1$. The distribution of $X_{i4}^{*(k)}$ given $\boldsymbol{\gamma}_{04}, \boldsymbol{\gamma}_{14}$ and the data, $\mathbf{X}, \mathbf{Z}$, is $\mathcal{N}\left(\gamma_{04}^{(k)} + \gamma_{14}^{(k)} Z_i, 1\right)$ truncated either to the left or to the right of zero depending on the observed value X_{i4}. We have

$$\gamma_{04}^{(k)} \mid \gamma_{14}, \sigma_{\gamma_0}^2, \mu_{\gamma_0}, \mathbf{X}, \mathbf{Z}, \mathbf{X}_4^* \sim N\left(\overline{\mu}_{\gamma_{04}}^{(k)}, \overline{\sigma}_{\gamma_{04}}^{2(k)}\right)$$

where $\mathbf{X}_4^* = \left(X_4^{*(1)}, \ldots, X_j^{*(K-1)}\right)$ and

$$\overline{\sigma}_{\gamma_{04}}^{2(1)} = \left(\frac{1}{\sigma_{\gamma_0}^2} + N\right)^{-1} \qquad \overline{\mu}_{\gamma_{04}}^{(1)} = \overline{\sigma}_{\gamma_{04}}^2 \left[\frac{\mu_{\gamma_0}}{\sigma_{\gamma_0}^2} + \textstyle\sum_i (X_{i4}^{*(k)} - \gamma_{14} Z_i)\right]$$

$$\overline{\sigma}_{\gamma_{04}}^{2(k)} = \left(\frac{1}{\sigma_{\gamma_0}^2} + \textstyle\sum_i \mathbb{I}(X_{i4} \notin \{1, \ldots, k-1\})\right)^{-1}$$

$$\overline{\mu}_{\gamma_{04}}^{(k)} = \overline{\sigma}_{\gamma_{04}}^2 \left[\frac{\mu_{\gamma_0}}{\sigma_{\gamma_0}^2} + \textstyle\sum_{i: X_{i4} \notin \{1, \ldots, k-1\}} (X_{i4}^{*(k)} - \gamma_{14} Z_i)\right],$$

and

$$p\left(\gamma_{14}^{(k)} \mid \boldsymbol{\gamma}_{04}, \pi_{\gamma_1}, \sigma_{\gamma_1}^2, \mu_{\gamma_1}, \mathbf{X}, \mathbf{Z}\right) \propto \left[1 - \Phi\left(\gamma_{04}^{(k)} + \gamma_{14}^{(k)} Z_i\right)\right]^{\mathbb{I}\{X_{i4} = k\}}$$
$$\prod_{\ell = k+1, \ldots, K} \left[\Phi\left(\gamma_{04}^{(k)} + \gamma_{14}^{(k)} Z_i\right)^{\mathbb{I}\{X_{i4} = \ell\}}\right] \times$$
$$\left[\pi_{\gamma_1} \mathbb{I}(\{\gamma_{14}^{(k)} = 0\}) + (1 - \pi_{\gamma_1}) \frac{1}{\sqrt{2\pi\sigma_{\gamma_1}^2}} \exp\left\{-\frac{1}{2\sigma_{\gamma_1}^2}\left(\gamma_{14}^{(k)} - \mu_{\gamma_1}\right)^2\right\}\right].$$

The MCMC algorithm we use to sample from the posterior distributions of the parameters simulates successively from the above full conditional distributions. The data augmentation method is used to impute at each step the latent variables underlying binary and categorical covariates. Simulations from the mixture distributions for the values of $\gamma_{11}, \gamma_{12}, \gamma_{13}, \boldsymbol{\gamma}_{14}, \gamma_{15}$, and $\boldsymbol{\gamma}_{16}$ use Metropolis-Hastings step (Berry and Berry, 2004). Specifically, we draw from a point mass at zero with probability 0.5 and a Normal distribution centred on the current value of the parameter. Let γ_{1j}^{cand} be the candidate value and $\gamma_{1j}^{(t)}$ the current value at iteration t. For simplicity we omit superscript (k) for $j = 4$. Define

$$r_j = \frac{p\left(\gamma_{1j}^{cand} \mid \boldsymbol{\gamma}_{0j}, \pi_{\gamma_1}, \sigma_{\gamma_1}^2, \mu_{\gamma_1}, \mathbf{X}, \mathbf{Z}\right)}{p\left(\gamma_{1j}^{(t)} \mid \boldsymbol{\gamma}_{0j}, \pi_{\gamma_1}, \sigma_{\gamma_1}^2, \mu_{\gamma_1}, \mathbf{X}, \mathbf{Z}\right)},$$

where $\boldsymbol{\gamma}_{0j} = \gamma_{0j}$ for $j = 1, 2, 3, 5, 6$. The candidate draw is accepted with probability

$$r = \begin{cases} r_j & \text{if } \gamma_{1j}^{cand} = \gamma_{1j}^{(t)} = 0 \\ \dfrac{1}{\sqrt{2\pi\sigma_{MH}^2}} \exp\left\{ -\dfrac{(\gamma_{1j}^{(t)})^2}{2\sigma_{MH}^2} \right\} r_j & \text{if } \gamma_{1j}^{cand} = 0 \text{ and } \gamma_{1j}^{(t)} \neq 0 \\ \dfrac{r_j}{\dfrac{1}{\sqrt{2\pi\sigma_{MH}^2}} \exp\left\{ -\dfrac{(\gamma_{1j}^{(cand)})^2}{2\sigma_{MH}^2} \right\}} & \text{if } \gamma_{1j}^{cand} \neq 0 \text{ and } \gamma_{1j}^{(t)} = 0 \\ r_j & \text{if } \gamma_{1j}^{cand} \neq 0 \text{ and } \gamma_{1j}^{(t)} \neq 0 \end{cases}$$

where σ_{MH}^2 is the variance of the Normal proposal distribution.

1.9.2 Bayesian Principal Stratification Inference: MCMC Algorithm

Under the model specification described in Section 1.6.2 we have:

$$\begin{array}{rcll} \pi_i(AE) & \equiv & Pr(\underline{E}_i = AE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) & = & Pr(E_i^*(AE) \leq 0 \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\ \pi_i(NE) & \equiv & Pr(\underline{E}_i = NE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) & = & Pr(E_i^*(AE) > 0,\, E_i^*(NE) \leq 0 \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\ \pi_i(OC) & \equiv & Pr(\underline{E}_i = OC \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) & = & 1 - Pr(E_i = NE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) - \\ & & & & Pr(E_i = AE \mid \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \end{array}$$

where

$$E_i^*(AE) = \alpha_{AE}^{(0)} + \tilde{S}_i \alpha_{AE}^{(S)} + \mathbf{X}_i' \boldsymbol{\alpha}_{AE}^{(X)} + \epsilon_{i,AE}, \quad \epsilon_{i,AE} \sim \mathcal{N}(0,1)$$
$$E_i^*(NE) = \alpha_{NE}^{(0)} + \tilde{S}_i \alpha_{NE}^{(S)} + \mathbf{X}_i' \boldsymbol{\alpha}_{NE}^{(X)} + \epsilon_{i,NE}, \quad \epsilon_{i,NE} \sim \mathcal{N}(0,1).$$

with $\epsilon_{i,AE}$ and $\epsilon_{i,NE}$ independent.

$$\begin{array}{rcl} \pi_i(00 \mid AE) & \equiv & Pr(\underline{C}_i = (0,0) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\ & = & Pr(C_i^*(0,0) \leq 0 \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\ \pi_i(01 \mid AE) & \equiv & Pr(\underline{C}_i = (0,1) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\ & = & Pr(C_i^*(0,0) > 0,\, C_i^*(0,1) \leq 0 \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\ \pi_i(11 \mid AE) & \equiv & Pr(\underline{C}_i = (1,1) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\ & = & Pr(C_i^*(0,0) > 0,\, C_i^*(0,1) > 0, C_i^*(1,1) \leq 0 \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}), \\ \pi_i(10 \mid AE) & \equiv & Pr(\underline{C}_i = (1,0) \mid \underline{E}_i = AE, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) \\ & = & 1 - \pi_i(00 \mid AE) - \pi_i(01 \mid AE) - \pi_i(11 \mid AE) \end{array}$$

where $C_i^*(\underline{c}) = \beta_{AE_{\underline{c}}}^{(0)} + \tilde{S}_i \beta_{AE_{\underline{c}}}^{(S)} + \mathbf{X}_i' \boldsymbol{\beta}_{AE_{\underline{c}}}^{(X)} + \epsilon_{i,AE_{\underline{c}}}$ with $\epsilon_{i,AE_{\underline{c}}} \sim \mathcal{N}(0,1)$, independently, for $\underline{c} = (0,0), (0,1), (1,1)$.

$$\pi_i(*0 \mid OC) \equiv Pr(\underline{C}_i = (*,0) \mid \underline{E}_i = OC, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta}) = Pr(C_i^*(*,0) > 0 \mid \underline{E}_i = OC, \tilde{S}_i, \mathbf{X}_i; \boldsymbol{\theta})$$

where $C_i^*(*,0) = \beta_{OC_{*0}}^{(0)} + \tilde{S}_i\beta_{OC_{*0}}^{(S)} + \mathbf{X}_i'\boldsymbol{\beta}_{OC_{*0}}^{(X)} + \epsilon_{i,OC_{*0}}$ with $\epsilon_{i,OC_{*0}} \sim \mathcal{N}(0,1)$.

For $(z,b) \in \{(0,AE_{11}),(0,AE_{01}),(1,AE_{11}),(1,AE_{10}),(1,OC_{*0})\}$:

$$f_{i,z}^{(UC)}(y \mid b) \equiv Pr\left(\widetilde{Y}_{UC_i}(z) = y \mid B_i = b, S_i, \mathbf{X}_i; \boldsymbol{\theta}\right) = \frac{1}{y!}\lambda_i(z,b)^y e^{-\lambda_i(z,b)}$$

where

$$\log(\lambda_i(z,b)) = \gamma_{z,b}^{(0)} + \tilde{S}_i\gamma_{z,b}^{(S)} + \mathbf{X}_i'\boldsymbol{\gamma}_{z,b}^{(X)},$$

For $(z,b) \in \{(0,AE_{11}),(1,AE_{11}),(1,AE_{01}),(0,AE_{10}),(1,OC_{*1})\}$, let $\widetilde{Y}_{i,G}(z) = \log(Y_{i,G}(z))$ and

$$f_{i,z}^{(G)}(\tilde{y} \mid \widetilde{Y}_{i,UC}(z), B_i = b, \tilde{S}_i, \mathbf{X}_i) = \frac{\frac{1}{\sqrt{2\pi\sigma_{z,b}^2}}\exp\left\{\frac{1}{2\sigma_{z,b}^2}\left[\tilde{y} - \tau_i(z,b)\right]^2\right\}}{\Phi\left(\frac{\log(b) - \tau_i(z,b)}{\sqrt{\sigma_{z,b}^2}}\right) - \Phi\left(\frac{\log(a) - \tau_i(z,b)}{\sqrt{\sigma_{z,b}^2}}\right)},$$

where

$$\tau_i(z,b) = \eta_{z,b}^{(0)} + \tilde{S}_i\eta_{z,b}^{(S)} + Y_{i,UC}(z)\eta_{z,b}^{(UC)} + \mathbf{X}_i'\boldsymbol{\eta}_{z,b}^{(X)},$$

and $a = \log(18 - 1/3)$ and $b = \log(30 + 1/3)$.

Recall that we impose the following prior equalities: $\boldsymbol{\gamma}^{(X)} \equiv \boldsymbol{\gamma}_{z,b}^{(X)}$ for $(z,b) \in \{(0,AE_{11}),(1,AE_{11}),(1,AE_{01}),(0,AE_{10}),(1,OC_{*0}),(1,OC_{*1})\}$; $\boldsymbol{\eta}^{(X)} \equiv \boldsymbol{\eta}_{z,b}^{(X)}$ for $(z,b) \in \{(0,AE_{11}),(1,AE_{11}),(1,AE_{01}),(0,AE_{10}),(1,OC_{*1})\}$; and $\sigma_{AE}^2 \equiv \sigma_{0,AE_{11}}^2 = \sigma_{1,AE_{11}}^2 = \sigma_{0,AE_{10}}^2 = \sigma_{1,AE_{01}}^2$.

Prior distribution

We assume that parameters are a *priori* independent, and we use multivariate normal prior distributions for the regression coefficients and Scaled-Inverse-χ^2 prior distributions for the variances of the truncated log-normal models. Formally, $\boldsymbol{\alpha}_{\underline{e}} \equiv [\alpha_{\underline{e}}^{(0)}, \alpha_{\underline{e}}^{(S)}, \boldsymbol{\alpha}_{\underline{e}}^{(X)}]$ $\sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\alpha}_{\underline{e}}}; \underline{\sigma}_{\boldsymbol{\alpha}_{\underline{e}}}^2\mathbf{I})$; $\underline{e} = AE, NE$; $\boldsymbol{\beta}_{AE_{\underline{c}}} \equiv [\beta_{AE_{\underline{c}}}^{(0)}, \beta_{AE_{\underline{c}}}^{(S)}, \boldsymbol{\beta}_{AE_{\underline{c}}}^{(X)}] \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\beta}_{AE_{\underline{c}}}}; \underline{\sigma}_{\boldsymbol{\beta}_{AE_{\underline{c}}}}^2\mathbf{I})$, $\underline{c} = (0,0)$, $(0,1)$, $(1,1)$; $\boldsymbol{\beta}_{OC_{*0}} \equiv [\beta_{OC_{*0}}^{(0)}, \beta_{OC_{*0}}^{(S)}, \boldsymbol{\beta}_{OC_{*0}}^{(X)}] \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\beta}_{OC_{*0}}}; \underline{\sigma}_{\boldsymbol{\beta}_{OC_{*0}}}^2\mathbf{I})$; $\boldsymbol{\gamma}_{z,b} \equiv [\gamma_{z,b}^{(0)}, \gamma_{z,b}^{(S)}] \sim$ $\sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\gamma}_{z,b}}; \underline{\sigma}_{\boldsymbol{\gamma}_{z,b}}^2\mathbf{I})$, $(z,b) \in \{(0,AE_{11}),(1,AE_{11}),(1,AE_{01}),(0,AE_{10}),(1,OC_{*1})\}$, and $\boldsymbol{\gamma}^{(\mathbf{X})} \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\gamma}^{(X)}}; \underline{\sigma}_{\boldsymbol{\gamma}^{(X)}}^2\mathbf{I})$; $\boldsymbol{\eta}_{z,b} \equiv [\eta_{z,b}^{(0)}, \eta_{z,b}^{(S)}, \eta^{(UC)}] \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\eta}_{z,b}}; \underline{\sigma}_{\boldsymbol{\eta}_{z,b}}^2\mathbf{I})$, $(z,g) \in \{(0,AE_{11}),(1,AE_{11}),(1,AE_{01}),(0,AE_{10}),(1,OC_{*0}),(1,OC_{*1})\}$, $\boldsymbol{\eta}^{(\mathbf{X})} \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\eta}^{(X)}}; \underline{\sigma}_{\boldsymbol{\eta}^{(X)}}^2\mathbf{I})$, $\sigma_{AE}^2 \sim$ $\sim Scale{-}inv{-}\chi^2(\nu,\tau^2)$ and $\sigma_{OC_{*1}}^2 \sim Scale{-}inv{-}\chi^2(\nu,\tau^2)$, where $\nu = 0.02$ and $\tau^2 = 1$. We specify weakly informative priors using (multivariate) normal prior distributions centered in zero and variances equal to 10: $\underline{\mu}_{\boldsymbol{\alpha}_{\underline{e}}}$, $\underline{\mu}_{\boldsymbol{\beta}_{AE_{\underline{c}}}}$ $\underline{\mu}_{\boldsymbol{\beta}_{OC_{*0}}}$ $\underline{\mu}_{\boldsymbol{\gamma}_{z,b}}$, $\underline{\mu}_{\boldsymbol{\gamma}^{(X)}}$ $\underline{\mu}_{\boldsymbol{\eta}_{z,b}}$, $\underline{\mu}_{\boldsymbol{\eta}^{(X)}}$ and $\underline{\mu}_{\eta^{(UC)}}$ are all null vectors, and $\underline{\sigma}_{\boldsymbol{\alpha}_{\underline{e}}}^2 = \underline{\sigma}_{\boldsymbol{\beta}_{AE_{\underline{c}}}}^2 = \underline{\sigma}_{\boldsymbol{\beta}_{OC_{*0}}}^2 = \underline{\sigma}_{\boldsymbol{\gamma}_{z,b}}^2 = \underline{\sigma}_{\boldsymbol{\gamma}^{(X)}}^2 = \underline{\sigma}_{\boldsymbol{\eta}_{z,b}}^2 = \underline{\sigma}_{\boldsymbol{\eta}^{(X)}}^2 = 10$.

Let

$$\begin{aligned}\boldsymbol{\theta} = (&\alpha^{(0)}_{AE}, \alpha^{(S)}_{AE}, \boldsymbol{\alpha}^{(\mathbf{X})}_{AE}, \alpha^{(0)}_{NE}, \alpha^{(S)}_{NE}, \boldsymbol{\alpha}^{(\mathbf{X})}_{NE}, \beta^{(0)}_{AE_{00}}, \beta^{(S)}_{AE_{00}}, \boldsymbol{\beta}^{(\mathbf{X})}_{AE_{00}}, \beta^{(0)}_{AE_{01}}, \beta^{(S)}_{AE_{01}}, \boldsymbol{\beta}^{(\mathbf{X})}_{AE_{01}},\\ &\beta^{(0)}_{AE_{11}}, \beta^{(S)}_{AE_{11}}, \boldsymbol{\beta}^{(\mathbf{X})}_{AE_{11}}, \beta^{(0)}_{OC_{*0}}, \beta^{(S)}_{OC_{*0}}, \boldsymbol{\beta}^{(\mathbf{X})}_{OC_{*0}},\\ &\gamma^{(0)}_{0,AE_{10}}, \gamma^{(S)}_{0,AE_{10}}, \gamma^{(0)}_{0,AE_{11}}, \gamma^{(S)}_{0,AE_{11}}, \gamma^{(0)}_{1,AE_{01}}, \gamma^{(S)}_{1,AE_{01}}, \gamma^{(0)}_{1,AE_{11}}, \gamma^{(S)}_{1,AE_{11}}, \gamma^{(0)}_{1,OC_{*1}}, \gamma^{(S)}_{1,OC_{*1}}, \boldsymbol{\gamma}^{(X)},\\ &\eta^{(0)}_{0,AE_{10}}, \eta^{(S)}_{0,AE_{10}}, \eta^{(UC)}_{0,AE_{10}}, \eta^{(0)}_{0,AE_{11}}, \eta^{(S)}_{0,AE_{11}}, \eta^{(UC)}_{0,AE_{11}}, \eta^{(0)}_{1,AE_{01}}, \eta^{(S)}_{1,AE_{01}}, \eta^{(UC)}_{1,AE_{01}},\\ &\eta^{(0)}_{1,AE_{11}}, \eta^{(S)}_{1,AE_{11}}, \eta^{(UC)}_{1,AE_{11}},\\ &\eta^{(0)}_{1,OC_{*1}}, \eta^{(S)}_{1,OC_{*1}}, \eta^{(UC)}_{1,OC_{*1}}, \boldsymbol{\eta}^{(X)}, \sigma^2_{AE}, \sigma^2_{OC})\end{aligned}$$

denote the vector of the parameters, and let $p(\boldsymbol{\theta})$ denote the prior distribution for the full parameter vector, which is the product of the above prior distributions under assumption of prior independence of the parameters.

Posterior distribution

The posterior distribution of the parameters is obtained from Markov chain Monte Carlo (MCMC) methods. The Markov chain algorithm that we adopt is based on a sequence of Metropolis-Hastings and Gibbs sampler steps. The algorithm uses the data augmentation (DA) method to impute at each step the missing principal stratum indicators, $B_i = (\underline{E}_i, \underline{C}_i)$, and to exploit the complete principal strata posterior distribution to update the parameters.

Specifically, let $\mathbf{Z} = [Z_i]_{i\in\mathcal{U}_{s_0}}$, $\mathbf{E}^{obs} = [E_i^{obs}]_{i\in\mathcal{U}_{s_0}}$, $\mathbf{C}^{obs} = [C_i^{obs}]_{i\in\mathcal{U}_{s_0}}$, $\mathbf{Y}^{obs}_{UC} = [Y^{obs}_{i,UC}]_{i\in\mathcal{U}_{s_0}}$, $\mathbf{Y}^{obs}_{G} = [Y^{obs}_{i,G}]_{i\in\mathcal{U}_{s_0}}$, $\mathbf{S} = [S_i]_{i\in\mathcal{U}_{s_0}}$ and $\mathbf{X} = [\mathbf{X}_i]_{i\in\mathcal{U}_{s_0}}$. We first obtain the joint posterior distribution of $(\boldsymbol{\theta}, \mathbf{E}^{mis}, \mathbf{C}^{mis})$ from a Gibbs sampler by iteratively sampling from

$$p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{E}^{obs}, \mathbf{C}^{obs}, \mathbf{E}^{mis}, \mathbf{C}^{mis}, \mathbf{Y}^{obs}_{UC}, \mathbf{Y}^{obs}_{G}, \mathbf{S}, \mathbf{X}; \mathcal{U}_{s_0}),$$

$$p(\mathbf{E}^{mis} \mid \mathbf{Z}, \mathbf{E}^{obs}, \mathbf{C}^{obs}, \mathbf{Y}^{obs}_{UC}, \mathbf{Y}^{obs}_{G}, \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}; \mathcal{U}_{s_0}),$$

and

$$p(\mathbf{C}^{mis}, \mid \mathbf{Z}, \mathbf{E}^{obs}, \mathbf{E}^{mis}, \mathbf{C}^{obs}, \mathbf{Y}^{obs}_{UC}, \mathbf{Y}^{obs}_{G}, \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}; \mathcal{U}_{s_0}).$$

This procedure, in turn, provides the marginal posterior distribution

$$p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{E}^{obs}, \mathbf{C}^{obs}, \mathbf{Y}^{obs}_{UC}, \mathbf{Y}^{obs}_{G}, \mathbf{S}, \mathbf{X}; \mathcal{U}_{s_0})$$

and thus the posterior of the causal estimands of interest (see the following section for computational details). The key to the posterior computation is the evaluation of the complete principal-strata posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{E}^{obs}, \mathbf{C}^{obs}, \mathbf{E}^{mis}, \mathbf{C}^{mis}, \mathbf{Y}^{obs}_{UC}, \mathbf{Y}^{obs}_{G}, \mathbf{S}, \mathbf{X}; \mathcal{U}_{s_0})$,

which has a simple form without any mixture component:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{E}^{obs}, \mathbf{C}^{obs}, \mathbf{E}^{mis}, \mathbf{C}^{mis}, \mathbf{Y}^{obs}_{UC}, \mathbf{Y}^{obs}_{G}, \mathbf{S}, \mathbf{X}; \mathcal{U}_{s_0}) \propto p(\boldsymbol{\theta}) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 0,1, B_i = NE} \pi_i(NE) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 0,1, B_i = OC_{*0}} \pi_i(OC) \cdot \pi_i(*0 \mid_{OC}) \times \prod_{i \in \mathcal{U}_{s_0}: Z_i = 0, B_i = OC_{*1}} \pi_i(OC) \cdot \pi_i(*1 \mid_{OC}) \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 0, B_i = AE_{00}} \pi_i(AE) \cdot \pi_i(00 \mid AE) \times \prod_{i \in \mathcal{U}_{s_0}: Z_i = 0, B_i = AE_{01}} \pi_i(AE) \cdot \pi_i(01 \mid AE) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 0, B_i = AE_{11}} \pi_i(AE) \cdot \pi_i(11 \mid AE) \cdot f^{(UC)}_{i,0}(\widetilde{Y}^{obs}_{i,UC} \mid AE_{11}) \cdot f^{(G)}_{i,0}(\widetilde{Y}^{obs}_{i,G} \mid Y_{i,UC}(0), AE_{11}) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 0, B_i = AE_{10}} \pi_i(AE) \cdot \pi_i(10 \mid AE) \cdot f^{(UC)}_{i,0}(\widetilde{Y}^{obs}_{i,UC} \mid AE_{10}) \cdot f^{(G)}_{i,0}(\widetilde{Y}^{obs}_{i,G} \mid Y_{i,UC}(0), AE_{10}) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 1, B_i = AE_{10}} \pi_i(AE) \cdot \pi_i(10 \mid AE) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 1, B_i = AE_{11}} \pi_i(AE) \cdot \pi_i(11 \mid AE) \cdot f^{(UC)}_{i,1}(\widetilde{Y}^{obs}_{i,UC} \mid AE_{11}) f^{(G)}_{i,1}(\widetilde{Y}^{obs}_{i,G} \mid Y_{i,UC}(1), AE_{11}) \times \\
\prod_{i \in \mathcal{U}_{s_0}: Z_i = 1, B_i = AE_{01}} \pi_i(AE) \cdot \pi_i(01 \mid AE) \cdot f^{(UC)}_{i,1}(\widetilde{Y}^{obs}_{i,UC} \mid AE_{01}) f^{(G)}_{i,1}(\widetilde{Y}^{obs}_{i,G} \mid Y_{i,UC}(1), AE_{01}).
\end{aligned}
$$

Details of calculations

Given $(\mathbf{B}^{(t)}, \boldsymbol{\theta}^{(t)}) = (\underline{\mathbf{E}}^{(t)}, \underline{\mathbf{C}}^{(t)}, \boldsymbol{\theta}^{(t)})$, the principal strata indicator and the state of the chain at time t, the state of the chain at time $t+1$ follows from applying the following steps.

1. Sample $\mathbf{B}^{(t+1)}$. Note that conditional on $\boldsymbol{\theta}$ and the observed data, $\mathbf{Z}$, $\mathbf{E}^{obs}$, $\mathbf{C}^{obs}$, $\mathbf{Y}^{obs}_{UC}$, $\mathbf{Y}^{obs}_{G}$, $\mathbf{X}$, $\tilde{\mathbf{S}}$, the B_i are independent. Moreover, we have:

$$
\begin{aligned}
& Pr(B_i = b \mid \boldsymbol{\theta}^{(t)}, Z_i, E^{obs}_i, C^{obs}_i, Y^{obs}_{i,UC}, Y^{obs}_{i,G}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \\
= \;& Pr(\underline{E}_i = \underline{e} \mid \boldsymbol{\theta}^{(t)}, Z_i, E^{obs}_i, C^{obs}_i, Y^{obs}_{i,UC}, Y^{obs}_{i,G}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \times \\
& Pr(\underline{C}_i = \underline{c} \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i = \underline{e}, E^{obs}_i, C^{obs}_i, Y^{obs}_{i,UC}, Y^{obs}_{i,G}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}).
\end{aligned}
$$

   Exploiting the above factorization we first sample $\underline{E}_i$. We have:

   - $Pr(\underline{E}_i = NE \mid \boldsymbol{\theta}^{(t)}, Z_i, E^{obs}_i, C^{obs}_i, Y^{obs}_{i,UC}, Y^{obs}_{i,G}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) = 1$ if $Z_i = 1, E^{obs}_i = 0$,
   - $Pr(\underline{E}_i = NE \mid \boldsymbol{\theta}^{(t)}, Z_i, E^{obs}_i, C^{obs}_i, Y^{obs}_{i,UC}, Y^{obs}_{i,G}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \propto \dfrac{\pi_i(NE)}{\pi_i(NE) + \pi_i(OC)}$ if $Z_i = 0, E^{obs}_i = 0$,

   - $Pr(\underline{E}_i = AE \mid \boldsymbol{\theta}^{(t)}, Z_i, E^{obs}_i, C^{obs}_i, Y^{obs}_{i,UC}, Y^{obs}_{i,G}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) = 1$, if $Z_i = 0, E^{obs}_i = 1$,
   - Finally for observations with $Z_i = 1, E^{obs}_i = 1$,

$$
\begin{aligned}
&Pr(\underline{E}_i = AE \mid \boldsymbol{\theta}^{(t)}, Z_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \\
&\quad \propto \; \pi_i(AE) \cdot [\pi_i(00 \mid AE) + \pi_i(10 \mid AE) \\
&\qquad +\pi_i(11 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{11}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{11}) \\
&\qquad +\pi_i(01 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{01}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{01})] \times \\
&\qquad \{\pi_i(AE) \cdot [\pi_i(00 \mid AE) + \pi_i(10 \mid AE) \\
&\qquad +\pi_i(11 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{11}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{11}) \\
&\qquad +\pi_i(01 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{01}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{01})] \\
&\qquad +\pi_i(OC) \cdot [\pi_i(*0 \mid OC) \\
&\qquad +\pi_i(*1 \mid OC) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid OC_{*1}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, OC_{*1})]\}^{-1}.
\end{aligned}
$$

Given $\underline{E}_i$ we then sample $\underline{C}_i$ with probability

- for observations with $Z_i = 0, 1, \underline{E}_i = NE, E_i^{obs} = 0$

$$Pr(\underline{C}_i = (*, *) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) = 1;$$

- for observations with $Z_i = 1, \underline{E}_i = OC, E_i^{obs} = 1$ and $C_i^{obs} = 0$

$$Pr(\underline{C}_i = (*, 0) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) = 1;$$

- for observations with $Z_i = 1, \underline{E}_i = OC, E_i^{obs} = 1$ and $C_i^{obs} = 1$

$$Pr(\underline{C}_i = (*, 1) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) = 1;$$

- for observations with $Z_i = 0, \underline{E}_i = OC, E_i^{obs} = 0$, we have

$$
\begin{aligned}
Pr(\underline{C}_i = (*, 0) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) &\propto \\
\propto \frac{\pi_i(*0 \mid OC)}{\pi_i(*0 \mid OC) + \pi_i(*1 \mid OC)}&;
\end{aligned}
$$

- for observations with $Z_i = 0, \underline{E}_i = AE, E_i^{obs} = 1$ and $C_i^{obs} = 0$, we have

$$
\begin{aligned}
Pr(\underline{C}_i = (0, 0) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) &\propto \\
\propto \frac{\pi_i(00 \mid AE)}{\pi_i(00 \mid AE) + \pi_i(01 \mid AE)}&;
\end{aligned}
$$

- for observations with $Z_i = 0, \underline{E}_i = AE, E_i^{obs} = 1$ and $C_i^{obs} = 1$ we have

$$\begin{aligned} &Pr(\underline{C}_i = (1,1) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \\ &\quad \propto \pi_i(11 \mid AE) \cdot f_{i,0}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{11}) \cdot f_{i,0}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{11}) \\ &\qquad \times [\pi_i(11 \mid AE) \cdot f_{i,0}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{11}) \cdot f_{i,0}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{11}) \\ &\qquad + \pi_i(10 \mid AE) \cdot f_{i,0}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{10}) \cdot f_{i,0}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{10})]^{-1}; \end{aligned}$$

- for observations with $Z_i = 1, \underline{E}_i = AE, E_i^{obs} = 1$ and $C_i^{obs} = 1$ we have

$$\begin{aligned} &Pr(\underline{C}_i = (1,1) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i = AE, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \\ &\quad \propto \pi_i(11 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{11}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{11}) \\ &\qquad \times [\pi_i(11 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{11}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{11}) \\ &\qquad + \pi_i(01 \mid AE) \cdot f_{i,1}^{UC}(\tilde{Y}_{i,UC}^{obs} \mid AE_{01}) \cdot f_{i,1}^{G}(\log(Y_{i,G}^{obs}) \mid \tilde{Y}_{i,UC}^{obs}, AE_{01})]^{-1}; \end{aligned}$$

- for observations with $Z_i = 1, \underline{E}_i = AE, E_i^{obs} = 1$ and $C_i^{obs} = 0$ we have

$$\begin{aligned} &Pr(\underline{C}_i = (0,0) \mid \boldsymbol{\theta}^{(t)}, Z_i, \underline{E}_i = AE, E_i^{obs}, C_i^{obs}, Y_{i,UC}^{obs}, Y_{i,G}^{obs}, \mathbf{X}_i, \tilde{S}_i; \mathcal{U}_{s_0}) \\ &\qquad \propto \frac{\pi_i(00 \mid AE)}{\pi_i(00 \mid AE) + \pi_i(10 \mid AE)}. \end{aligned}$$

2. Sample the latent variables $E_i^*(AE)$ and $E_i^*(NE)$:

   (a) Sample the latent variable $E_i^*(AE)$ from $\mathcal{N}(\alpha_{AE}^{(0)^{(t)}} + \alpha_{AE}^{(S)^{(t)}} \tilde{S}_i + \boldsymbol{\alpha}_{AE}^{(\mathbf{X})^{(t)}} \mathbf{X}_i, 1)$ truncated to $(-\infty, 0)$ if $\underline{E}_i = AE$ and to $(0, \infty)$ if $\underline{E}_i \neq AE$.

   (b) Sample the latent variable $E_i^*(NE)$ from$\mathcal{N}(\alpha_{NE}^{(0)^{(t)}} + \alpha_{NE}^{(S)^{(t)}} \tilde{S}_i + \boldsymbol{\alpha}_{NE}^{(\mathbf{X})^{(t)}} \mathbf{X}_i, 1)$ truncated to $(-\infty, 0)$ if $\underline{E}_i = NE$ and to $(0, \infty)$ if $\underline{E}_i \neq NE$.

3. Sample the coefficients $\boldsymbol{\alpha}_{\underline{e}}^{(t+1)} \equiv [\alpha_{\underline{e}}^{(0)^{(t+1)}}, \alpha_{\underline{e}}^{(S)^{(t+1)}}, \boldsymbol{\alpha}_{\underline{e}}^{(X)^{(t+1)}}]$ from $\mathcal{N}(\mu_{\boldsymbol{\alpha}_{\underline{e}}}; \Sigma_{\boldsymbol{\alpha}_{\underline{e}}})$, for $\underline{e} = AE, NE$.

   Let $\mathbf{SX} = \left[\mathbf{1} | \tilde{\mathbf{S}} | \mathbf{X}\right]$ denote the $N \times (p+2)$ matrix with i^{th} row equals to $(1, \tilde{S}_i, X_{i1}, \ldots, X_{ip})$.

   (3.a) Sample $\boldsymbol{\alpha}_{AE}^{(t+1)}$ from $\mathcal{N}(\mu_{\alpha_{AE}}, \Sigma_{\alpha_{AE}})$ where

$$\mu_{\alpha_{AE}} = \Sigma_{\alpha_{AE}} \left((\underline{\sigma}_{\alpha_{AE}}^2 \mathbf{I})^{-1} \underline{\mu}_{\alpha_{AE}} + \mathbf{SX}' \mathbf{E}^*(AE) \right)$$

and

$$\Sigma_{\alpha_{AE}} = \left((\underline{\sigma}_{\alpha_{AE}}^2 \mathbf{I})^{-1} + \mathbf{SX}' \mathbf{SX} \right)^{-1}$$

   (3.b) Let $\mathbf{SX}_{OC,NE}$ denote the sub-matrix of $\mathbf{SX}$ with $\underline{E}_i^{(t+1)} = OC$ or $\underline{E}_i^{(t+1)} = NE$ and let $\mathbf{E}_{NE,OC}^*(NE)$ be the sub-vector of $\mathbf{E}^*(NE)$ for units with $\underline{E}_i^{(t+1)} = OC$ or $\underline{E}_i^{(t+1)} = NE$.

Sample $\boldsymbol{\alpha}_{NE}^{(t+1)}$ from $\mathcal{N}(\mu_{\alpha_{NE}}, \Sigma_{\alpha_{NE}})$ where

$$\mu_{\alpha_{NE}} = \Sigma_{\alpha_{NE}} \left((\underline{\sigma}^2_{\alpha_{NE}} \mathbf{I})^{-1} \underline{\mu}_{\alpha_{NE}} + \mathbf{SX}'_{OC,NE} \mathbf{E}^*_{OC,NE}(NE) \right)$$

and

$$\Sigma_{\alpha_{NE}} = \left((\underline{\sigma}^2_{\alpha_{NE}} \mathbf{I})^{-1} + \mathbf{SX}'_{OC,NE} \mathbf{SX}_{OC,NE} \right)^{-1}$$

4. Sample the latent variables C_i^* given $\underline{E}_i = AE$, which we denote with $C_i^*(\underline{c}) \mid_{AE}$, from $\mathcal{N}(\beta_{AE_{\underline{c}}}^{(0)^{(t)}} + \beta_{AE_{\underline{c}}}^{(S)^{(t)}} \tilde{S}_i + \boldsymbol{\beta}_{AE_{\underline{c}}}^{(\mathbf{X})^{(t)}} \mathbf{X}_i, 1)$ truncated to $(-\infty, 0)$ if $\underline{C}_i = \underline{c}$, and to $(0, \infty)$ if $\underline{C}_i \neq \underline{c}$, for $\underline{c} \in \{(0,0), (0,1), (1,1)\}$.

5. Sample the coefficients:
$\boldsymbol{\beta}_{AE_{\underline{c}}}^{(t+1)} \equiv [\boldsymbol{\beta}_{AE_{\underline{c}}}^{(0)^{(t+1)}}, \boldsymbol{\beta}_{AE_{\underline{c}}}^{(S)^{(t+1)}}, \boldsymbol{\beta}_{AE_{\underline{c}}}^{(X)^{(t+1)}}]$ from $\mathcal{N}(\mu_{\boldsymbol{\beta}_{AE_{\underline{c}}}}, \Sigma_{\boldsymbol{\beta}_{AE_{\underline{c}}}})$, for $\underline{c} \in \{(0,0), (0,1), (1,1)\}$, given $\underline{E}_i = AE$

   5.a Sample $\boldsymbol{\beta}_{AE_{00}}^{(t+1)}$ from $\mathcal{N}(\mu_{\boldsymbol{\beta}_{AE_{00}}}, \Sigma_{\boldsymbol{\beta}_{AE_{00}}})$ where

$$\mu_{\boldsymbol{\beta}_{AE_{00}}} = \Sigma_{\boldsymbol{\beta}_{AE_{00}}} \left((\underline{\sigma}^2_{\boldsymbol{\beta}_{AE_{00}}} \mathbf{I})^{-1} \underline{\mu}_{\boldsymbol{\beta}_{AE_{00}}} + \mathbf{SX}' \mathbf{C}^*(0,0) \mid_{AE} \right)$$

   and

$$\Sigma_{\boldsymbol{\beta}_{AE_{00}}} = \left((\underline{\sigma}^2_{\boldsymbol{\beta}_{AE_{00}}} \mathbf{I})^{-1} + \mathbf{SX}' \mathbf{SX} \right)^{-1}$$

   5.b Let $\mathbf{SX}_{AE_{01},AE_{10},AE_{11}}$ denote the sub-matrix of $\mathbf{SX}$ with $\underline{C}_i^{(t+1)} = (0,1) \mid_{AE}$ or $\underline{C}_i^{(t+1)} = (1,0) \mid_{AE}$ or $\underline{C}_i^{(t+1)} = (1,1) \mid_{AE}$ and let $\mathbf{C}^*_{AE_{01},AE10,AE_{11}}(0,1) \mid_{AE}$ be the sub-vector of $\mathbf{C}^*(0,1) \mid_{AE}$ for units with $\underline{C}_i^{(t+1)} = (0,1) \mid_{AE}$ or $\underline{C}_i^{(t+1)} = (1,0) \mid_{AE}$ or $\underline{C}_i^{(t+1)} = (1,1) \mid_{AE}$.
   Sample $\boldsymbol{\beta}_{AE_{01}}^{(t+1)}$ from $\mathcal{N}(\mu_{\boldsymbol{\beta}_{AE_{01}}}, \Sigma_{\boldsymbol{\beta}_{AE_{01}}})$ where

$$\mu_{\boldsymbol{\beta}_{AE_{01}}} = \Sigma_{\boldsymbol{\beta}_{AE_{01}}} \left((\underline{\sigma}^2_{\boldsymbol{\beta}_{AE_{01}}} \mathbf{I}^{-1}) \underline{\mu}_{\boldsymbol{\beta}_{AE_{01}}} + \mathbf{SX}'_{AE_{01},AE_{10},AE_{11}} \mathbf{C}^*_{AE_{01},AE10,AE_{11}}(0,1) \mid_{AE} \right)$$

   5.c Let $\mathbf{SX}_{AE_{10},AE_{11}}$ denote the sub-matrix of $\mathbf{SX}$ with $\underline{C}_i^{(t+1)} = (1,0) \mid_{AE}$ or $\underline{C}_i^{(t+1)} = (1,1) \mid_{AE}$
   and let $\mathbf{C}^*_{AE_{10},AE_{11}}(1,1) \mid_{AE}$ be the sub-vector of $\mathbf{C}^*(1,1) \mid_{AE}$ for units with $\underline{C}_i^{(t+1)} = (1,0) \mid_{AE}$ or $\underline{C}_i^{(t+1)} = (1,1) \mid_{AE}$.
   Sample $\boldsymbol{\beta}_{AE_{11}}^{(t+1)}$ from $\mathcal{N}(\mu_{\boldsymbol{\beta}_{AE_{11}}}, \Sigma_{\boldsymbol{\beta}_{AE_{11}}})$ where

$$\mu_{\boldsymbol{\beta}_{AE_{11}}} = \Sigma_{\boldsymbol{\beta}_{AE_{11}}} \left((\underline{\sigma}^2_{\boldsymbol{\beta}_{AE_{11}}} \mathbf{I})^{-1} \underline{\mu}_{\boldsymbol{\beta}_{AE_{11}}} + \mathbf{SX}'_{AE_{10},AE_{11}} \mathbf{C}^*_{AE_{10},AE_{11}}(1,1) \mid_{AE} \right)$$

   and

$$\Sigma_{\beta_{AE_{11}}} = \left((\underline{\sigma}^2_{\beta_{AE_{11}}} \mathbf{I}^{-1}) + \mathbf{SX}'_{AE10,AE_{11}} \mathbf{SX}_{AE_{10},AE_{11}} \right)^{-1}$$

6. Sample the latent variables C_i^* given $\underline{E}_i = OC$, e.g. $C_i^*(\underline{c}) \mid_{OC}$ from: $\mathcal{N}(\beta_{OC_{*0}}^{(0)^{(t)}} + \beta_{OC_{*0}}^{(S)^{(t)}} \tilde{S}_i + \boldsymbol{\beta}_{OC_{*0}}^{(\mathbf{X})^{(t)}} \mathbf{X}_i, 1)$ truncated to $(-\infty, 0)$ if $\underline{C}_i = (*, 0)$ and to $(0, \infty)$ if $\underline{C}_i = (*, 1)$.

7. Sample the coefficients: $\boldsymbol{\beta}_{OC_{*0}}^{(t)} \equiv [\beta_{OC_{*0}}^{(0)^{(t+1)}}, \beta_{OC_{*0}}^{(S)^{(t+1)}}, \boldsymbol{\beta}_{OC_{*0}}^{(X)^{(t+1)}}]$ from $\mathcal{N}(\mu_{\boldsymbol{\beta}_{OC_{*0}}}, \Sigma_{\boldsymbol{\beta}_{OC_{*0}}})$, for $\underline{C}_i = (*, 0)$, given $\underline{E}_i = OC$. Where

$$\mu_{\boldsymbol{\beta}_{OC_{*0}}} = \Sigma_{\boldsymbol{\beta}_{OC_{*0}}} \left((\underline{\sigma}^2_{\boldsymbol{\beta}_{OC_{*0}}} \mathbf{I})^{-1} \underline{\mu}_{\boldsymbol{\beta}_{OC_{*0}}} + \mathbf{SX'C}^*(0,0) \mid_{AE} \right)$$

and

$$\Sigma_{\boldsymbol{\beta}_{OC_{*0}}} = \left((\underline{\sigma}^2_{\boldsymbol{\beta}_{OC_{*0}}} \mathbf{I})^{-1} + \mathbf{SX'SX} \right)^{-1}$$

For each $i \in \mathcal{U}_{s_0}$ with $B_i = b$ and $Z_i \equiv \mathbb{I}\{S_i \geq s_0\} = z$, we draw the rest of the coefficients involved in the distributions of the two outcomes, $\widetilde{Y}_{UC}$ and $\widetilde{Y}_G$, in sequence as follows:

8. Sample the coefficients for the distribution of outcome $\widetilde{Y}_{UC}$: $\gamma_{0,AE_{10}}^{(0)^{(t+1)}}$, $\gamma_{0,AE_{10}}^{(S)^{(t+1)}}$, $\gamma_{0,AE_{11}}^{(0)^{(t+1)}}$, $\gamma_{0,AE_{11}}^{(S)^{(t+1)}}$, $\gamma_{1,AE_{01}}^{(0)^{(t+1)}}$, $\gamma_{1,AE_{01}}^{(S)^{(t+1)}}$, $\gamma_{1,AE_{11}}^{(0)^{(t+1)}}$, $\gamma_{1,AE_{11}}^{(S)^{(t+1)}}$, $\gamma_{1,OC_{*1}}^{(0)^{(t+1)}}$, $\gamma_{1,OC_{*1}}^{(S)^{(t+1)}}$, $\boldsymbol{\gamma}^{(X)^{(t+1)}}$ in sequence, conditional on all the others.
   In our specification, we could not draw directly from the appropriate full conditional distributions of the parameters related to the outcome distribution of University Credits; therefore we calculate the complete-data posterior density up to a normalizing constant at any parameter value, and use Metropolis-Hastings steps.
   Specifically, to draw $\gamma_{0,AE_{10}}^{(0)^{(t+1)}}$, we:

   8.a draw candidate value γ^{cand} from $\mathcal{N}(\gamma_{0,AE_{10}}^{(0)^{(t)}}, \sigma^2_{MH})$, where $\gamma_{0,AE_{10}}^{(0)^{(t)}}$ is the value of the parameter at time t, σ^2_{MH} is the variance of the Normal proposal distribution;

   8.b evaluate the complete-data posterior density of $\boldsymbol{\theta}$ at parameter vector $\boldsymbol{\theta}_{\gamma^{cand}}^{(t)}$, that is the vector of the parameters' values at iteration t replacing the value of the parameter $\gamma_{0,AE_{10}}^{(0)^{(t)}}$ with γ^{cand};

   8.c evaluate the complete-data posterior density of $\boldsymbol{\theta}$ at parameter vector $\boldsymbol{\theta}^{(t)}$;

   8.d sample a value from a Uniform distribution, $U(0, 1)$

   8.e accept the candidate draw if the ratio between the value at step 8.b and the value at step 8.c is strictly less than the value sampled at step 8.d.

   We then repeat the steps 8.a-8.e in sequence for parameters

$$\gamma_{0,AE_{10}}^{(S)}, \gamma_{0,AE_{11}}^{(0)}, \gamma_{0,AE_{11}}^{(S)}, \gamma_{1,AE_{01}}^{(0)}, \gamma_{1,AE_{01}}^{(S)}, \gamma_{1,AE_{11}}^{(0)}, \gamma_{1,AE_{11}}^{(S)}, \gamma_{1,OC_{*1}}^{(0)}, \gamma_{1,OC_{*1}}^{(S)}, \boldsymbol{\gamma}^{(X)}.$$

9. Sample the coefficients involved in the mean of the distribution of the outcome $\widetilde{Y}_G$, for which a truncated log-normal distribution to the interval $[\log(18-1/3), \log(30+1/3)]$ is specified.
   Also here in a similar way done in step 8, we calculate the complete-data posterior density up to a normalizing constant at any parameter value, and use Metropolis-Hastings steps in sequence for the parameters: $\eta_{0,AE_{10}}^{(0)^{(t+1)}}$, $\eta_{0,AE_{10}}^{(S)^{(t+1)}}$, $\eta_{0,AE_{10}}^{(UC)^{(t+1)}}$, $\eta_{0,AE_{11}}^{(0)^{(t+1)}}$, $\eta_{0,AE_{11}}^{(S)^{(t+1)}}$, $\eta_{0,AE_{11}}^{(UC)^{(t+1)}}$, $\eta_{1,AE_{01}}^{(0)^{(t+1)}}$, $\eta_{1,AE_{01}}^{(S)^{(t+1)}}$, $\eta_{1,AE_{01}}^{(UC)^{(t+1)}}$, $\eta_{1,AE_{11}}^{(0)^{(t+1)}}$, $\eta_{1,AE_{11}}^{(S)^{(t+1)}}$, $\eta_{1,AE_{11}}^{(UC)^{(t+1)}}$, $\eta_{1,OC_{*1}}^{(0)^{(t+1)}}$, $\eta_{1,OC_{*1}}^{(S)^{(t+1)}}$, $\eta_{1,OC_{*1}}^{(UC)^{(t+1)}}$, $\boldsymbol{\eta}^{(X)^{(t+1)}}$, $\sigma_{AE}^{2^{(t+1)}}$.

10. Sample values for parameters $\sigma_{AE}^{2^{(t+1)}}$, $\sigma_{1,OC}^{2^{(t+1)}}$, which are the variances of the outcome distribution $\widetilde{Y}_G$ for stratum AE and OC respectively.
    Specifically, to draw $\sigma_{AE}^{2^{(t+1)}}$ we use the proposal defined by a truncated normal distribution, truncated in $[a, b] = [0, +\infty)$, $\mathcal{TN}_{[a,b]}(\mu, \sigma)$, we proceed with the following steps:

    10.a draw candidate value $\sigma_{AE}^{2,cand}$ from $\mathcal{TN}_{[a,b]}(\sigma_{AE}^{2^{(t)}} - \dfrac{\phi(a)-\phi(b)}{\Phi(b)-\Phi(a)} \cdot \sigma_{MH}, \sigma_{MH}^2)$, where $\sigma_{AE}^{2^{(t)}}$ is the value of the variance for strata AE at time t, σ_{MH}^2 is the variance of the truncated normal proposal distribution, $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is the related cumulative distribution function;

    10.b evaluate $c.d.p.d_{cand}$ the complete-data posterior density of $\boldsymbol{\theta}$ at parameter vector $\boldsymbol{\theta}_{\sigma_{AE}^{2,cand}}^{(t)}$, that is the vector of the parameters' values at iteration t replacing the value of the parameter $\sigma_{AE}^{2^{(t)}}$ with $\sigma_{AE}^{2,cand}$;

    10.c evaluate $c.d.p.d_{old}$ the complete-data posterior density of $\boldsymbol{\theta}$ at parameter vector $\boldsymbol{\theta}^{(t)}$;

    10.d sample a value from a Uniform distribution, $U(0, 1)$

    10.e accept the candidate draw $\sigma_{AE}^{2,cand}$ if the value of the following ratio:

$$rr = \frac{c.d.p.d_{cand}}{c.d.p.d_{old}} \cdot \frac{d\mathcal{TN}_{[a,b]}(\sigma_{AE}^{2^{(t)}}, \mu = \sigma_{AE}^{2,cand} - \omega \cdot \sigma_{MH}, \sigma_{MH}^2)}{d\mathcal{TN}_{[a,b]}(\sigma_{AE}^{2,cand}, \mu = \sigma_{AE}^{2^{(t)}} - \omega \cdot \sigma_{MH}, \sigma_{MH}^2)}$$

    is strictly less than the value sampled at step 10.d. Where $\omega = \dfrac{\phi(a)-\phi(b)}{\Phi(b)-\Phi(a)}$, $d\mathcal{TN}_{[a,b]}$ is the density function of the truncated normal distribution in [a,b].

    We repeat steps 10.a-10.e also for $\sigma_{1,OC}^{2^{(t+1)}}$.

1.9.3 Sensitivity analyses

Sensitivity analysis w.r.t. Assumption 3 (Monotonicity of enrolment)

Assumption 3 of Monotonicity of enrolment allows to rule out the existence of OFA defiers principal stratum. In order to show the plausibility of this assumption we conduct a sensitivity analysis w.r.t. Assumption 3. In Bayesian analysis we can assess the sensitivity of the results to the monotonicity assumption by examining how the posterior distributions for the causal estimands change. Given the reduced sample size we conduct a sensitivity analysis with respect to the monotonicity assumption using a simplified setting where focus is on assess causal effects of OFA on the career indicator, C. Table 1.6 and Figure 1.3 respectively show posterior medians and 95% posterior credible intervals for the causal estimands of interest and the posterior distribution of the Survivor Average Causal Effects (SACEs) on students' academic career, $SACE_{\mathcal{U}_{s_0}}(C)$, with and without the monotonicity assumption. Results appear to be robust with respect to the monotonicity assumption. Both analyses suggest that most students are either always-enrolled or never-enrolled students: there are more than 56% always-enrolled students and more than 43% of never-enrolled students. The posterior median of the proportion of OFA compliers is approximately zero. The probability that there are OFA compliers is about 34.2% without monotonicity and 47.9% with monotonicity. In absence of monotonicity the posterior median of the proportion of OFA defiers is also approximately zero and probability that there are OFA defiers is less than 5%. The estimate of the Survivor average causal effect is approximately equal to 5.4%, with a standard deviation of 0.059, in absence of monotonicity and to 5.5%, with a standard deviation of 0.067 under monotonicity. In both scenarios the posterior probability that SACE is positive, that is, that not receiving OFA increases the probability of being an active student, is approximately 80% and the posterior distributions of SACE with and without monotonicity look very similar.

Sensitivity analysis w.r.t. prior specifications for σ^2_{AE} and $\sigma^2_{1,OC}$

We assess the sensitivity of the results with respect to the Scaled-Inverse-χ^2 prior distributions for the parameters σ^2_{AE} and $\sigma^2_{1,OC}$ of the truncated log-normal distributions of $\widetilde{Y}_G$ by repeating the analysis using the following uninformative prior distributions: $p(\sigma^2_{AE}) \propto 1/\sigma^2_{AE}$ and $p(\sigma^2_{1,OC}) \propto 1/\sigma^2_{1,OC}$ Tables 1.7 and 1.8 show the posterior medians and 95% posterior credible intervals (PCIs) of the posterior distributions of the principal strata proportions and of the causal estimands under monotonicity (Assumption 3) obtained using these priors. Results obtained are in line with those obtained using the Scaled-Inverse-χ^2 prior distribution for the variance parameter, $\sigma^2_{\underline{e}}$. The comparison between Tables 4 and 5 in the main text and Tables 1.7 and 1.8 suggests that our findings are robust with respect to the specification of the prior distribution for the parameters σ^2_{AE} and $\sigma^2_{1,OC}$: the results change only slightly under the two alternative prior specifications.

Table 1.1: Posterior probabilities of zero mean differences between groups of test participants defined by the OFA status.

	Bandwidth				
	$h = 2.0$	$h = 2.5$	$h = 3.0$	$h = 3.5$	$h = 4.0$
Covariate (sample size)	$(n = 297)$	$(n = 365)$	$(n = 434)$	$(n = 498)$	$(n = 571)$
Gender	0.901	0.886	0.902	0.901	0.898
High school Year (2011 vs <2011)	0.301	0.378	0.205	0.274	0.393
Irregular career	0.871	0.898	0.872	0.893	0.905
High school type (Baseline: Other)					
Humanity	0.884	0.867	0.849	0.883	0.898
Science	0.454	0.375	0.116	0.163	0.100
Tech	0.884	0.867	0.849	0.883	0.898
High school Grade	0.908	0.926	0.892	0.880	0.796
Far-away resident	0.890	0.885	0.871	0.897	0.890

Table 1.2: Means for the sub-populations $\mathcal{U}_{s_0}$, of the observed post treatment variables E, C, Y_{UC} and Y_G grouped by the OFA status Z_i.

Variables	$Z_i = 1$	$Z_i = 0$
E^{obs}	0.613	0.514
C^{obs}	0.714	0.625
Y_{UC}^{obs}	22.588	19.159
Y_G^{obs}	22.322	21.869

Table 1.3: Observed strata defined by the observed OFA status, Z, the observed status, E^{obs}, and the observed academic career status, C^{obs}, and corresponding latent principal strata (under Assumption 1.3). Symbol $*$ denotes that post treatment variable is truncated by the career status.

$(Z_i, E_i^{obs}, C_i^{obs})$	$Y_{i,UC}^{obs}$	$Y_{i,G}^{obs}$	Principal Strata
$(0,0,*)$	$*$	$*$	OC_{*0} or OC_{*1} or NE_{**}
$(0,1,0)$	0	$*$	AE_{00} or AE_{01}
$(0,1,1)$	>0	$\in [18,30]$	AE_{11} or AE_{10}
$(1,0,*)$	$*$	$*$	NE_{**}
$(1,1,0)$	0	$*$	AE_{00} or AE_{10} or OC_{*0}
$(1,1,1)$	>0	$\in [18,30]$	AE_{11} or AE_{01} or OC_{*1}

Table 1.4: Posterior median and 95% Posterior Credible Interval (PCI) based on the quantiles of the posterior distributions of the proportions of principal strata defined by $\underline{E}_i$ and by $(\underline{E}_i, \underline{C}_i)$ (left panel); and the proportions of principal strata defined by $\underline{C}_i$ among always-enrolled students ($\underline{E}_i = AE$) and among OFA compliers ($\underline{E}_i = OC$) (right panel).

Principal Stratum	50%	2.5%	97.5%	Principal Stratum	50%	2.5%	97.5%
NE_{**}	0.419	0.392	0.430				
AE	0.471	0.315	0.545				
AE_{00}	0.055	0.022	0.112	$00 \mid AE$	0.122	0.044	0.246
AE_{01}	0.121	0.063	0.156	$01 \mid AE$	0.257	0.176	0.321
AE_{10}	0.088	0.033	0.123	$10 \mid AE$	0.188	0.085	0.246
AE_{11}	0.200	0.134	0.249	$11 \mid AE$	0.431	0.376	0.490
OC	0.110	0.030	0.282				
OC_{*0}	0.033	0.011	0.079	$*0 \mid OC$	0.289	0.184	0.458
OC_{*1}	0.077	0.019	0.205	$*1 \mid OC$	0.711	0.542	0.816

Table 1.5: Posterior median and 95% Posterior credible interval (PCI) based on the quantiles of the posterior distribution of survivor average causal effects of OFA on: Enrolment (E), Career status (C), University Credits (UC) and average Grade (G).

Causal effect	50%	2.5%	97.5%
$ACE_{\mathcal{U}_{s_0}}(E)$	0.110	0.030	0.282
$SACE_{\mathcal{U}_{s_0}}(C)$	0.033	-0.003	0.071
$SACE_{\mathcal{U}_{s_0}}(UC)$	1.634	-7.378	10.614
$SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$	3.137	-14.848	22.414
$SACE_{\mathcal{U}_{s_0}}(G)$	0.410	-2.000	2.420

Figure 1.1: Histogram of the posterior distributions of $SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$, $SACE_{\mathcal{U}_{s_0}}(G)$ and $SACE_{\mathcal{U}_{s_0}}(UC)$ and corresponding posterior densities (derived using a kernel smoothing).

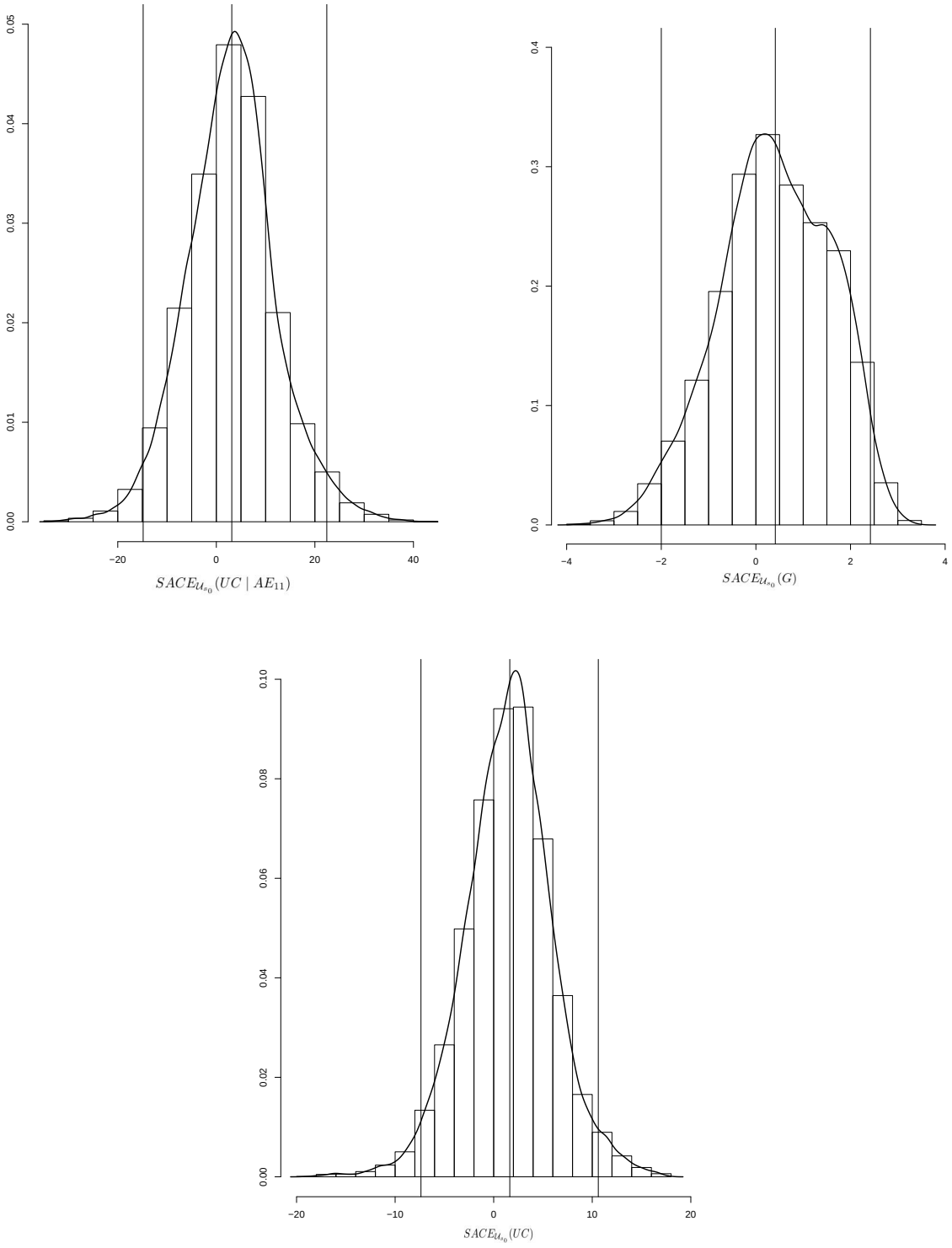


Figure 1.2: Simulation scatterplot of the joint posterior distribution of $SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$ and $SACE_{\mathcal{U}_{s_0}}(G)$.

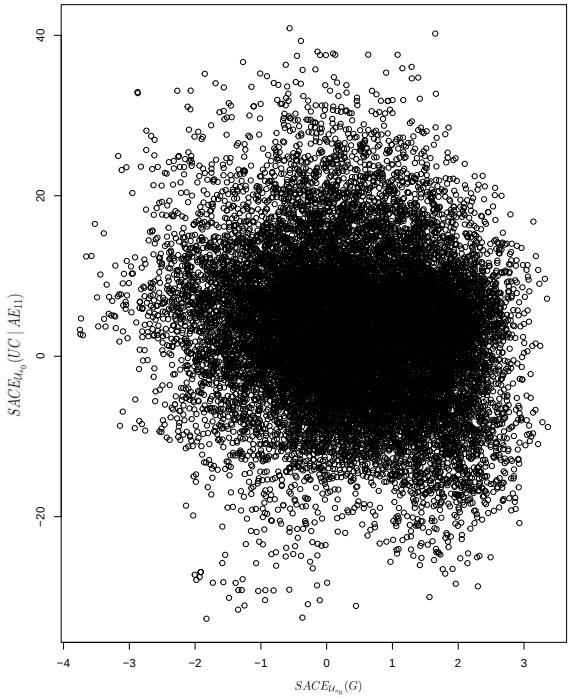


Table 1.6: Posterior median, 95% Posterior Credible Interval (PCI) based on the quantiles of the posterior distributions of the proportions of principal strata defined by $\underline{E}_i$ and of $SACE_{\mathcal{U}_{s_0}}(C)$: without monotonicity (left panel) and with monotonicity (right panel)

	Without Monotonicity			With Monotonicity		
Principal Stratum	50%	2.5%	97.5%	50%	2.5%	97.5%
AE	0.567	0.490	0.567	0.567	0.485	0.567
NE	0.433	0.373	0.433	0.433	0.370	0.433
OC	0.000	0.000	0.132	0.000	0.000	0.142
OD	0.000	0.000	0.008			
$SACE_{\mathcal{U}_{s_0}}(C)$	0.053	-0.065	0.175	0.053	-0.069	0.213

Figure 1.3: Posterior distributions of the Survivor Average Causal Effect on students' academic career ($SACE_{\mathcal{U}_{s_0}}(C)$) with (solid line) and without (dashed line) the monotonicity assumption.

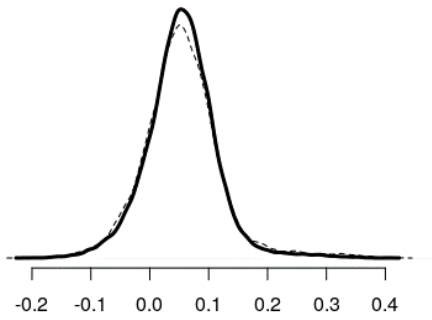


Table 1.7: Results of sensitivity analysis w.r.t. prior specification for σ_{AE} and $\sigma_{OC_{*1}}$. Posterior median and 95% Posterior Credible Interval (PCI) based on the quantiles of the posterior distributions of the proportions of principal strata defined by $\underline{E}_i$ and by $(\underline{E}_i, \underline{C}_i)$ (left panel); and the proportions of principal strata defined by $\underline{C}_i$ among always-enrolled students ($\underline{E}_i = AE$) and among OFA compliers ($\underline{E}_i = OC$) (right panel).

Principal Stratum	50%	2.5%	97.5%	Principal Stratum	50%	2.5%	97.5%
NE_{**}	0.419	0.392	0.433				
AE	0.479	0.323	0.556				
AE_{00}	0.052	0.019	0.112	$00 \mid AE$	0.111	0.041	0.241
AE_{01}	0.126	0.071	0.162	$01 \mid AE$	0.265	0.183	0.325
AE_{10}	0.090	0.044	0.129	$10 \mid AE$	0.192	0.105	0.246
AE_{11}	0.205	0.134	0.252	$11 \mid AE$	0.432	0.378	0.48
OC	0.104	0.019	0.271				
OC_{*0}	0.030	0.008	0.077	$*0 \mid OC$	0.291	0.143	0.500
OC_{*1}	0.071	0.014	0.200	$*1 \mid OC$	0.709	0.500	0.857

Table 1.8: Results of sensitivity analysis w.r.t. prior specification for σ_{AE} and $\sigma_{OC_{*1}}$. Posterior median and 95% Posterior credible interval (PCI) based on the quantiles of the posterior distribution of survivor average causal effects of OFA on: Enrolment (E), Career status (C), University Credits (UC) and average Grade (G).

Causal effect	50%	2.5%	97.5%
$ACE_{\mathcal{U}_{s_0}}(E)$	0.104	0.019	0.271
$SACE_{\mathcal{U}_{s_0}}(C)$	0.036	-0.003	0.071
$SACE_{\mathcal{U}_{s_0}}(UC)$	0.292	-9.365	8.938
$SACE_{\mathcal{U}_{s_0}}(UC \mid AE_{11})$	0.714	-17.359	19.548
$SACE_{\mathcal{U}_{s_0}}(G)$	0.713	-1.742	2.834

Chapter 2

The role of prior knowledge and numerical information in communicating epidemiological results to the population
A Bayesian Principal Stratification approach

Key words: Epidemiological results, Risk index, Causal effect, Principal Stratification, Bayesian Inference.

2.1 Introduction

The communication of the risk is important in every field: from marketing to medicine. The format used to communicate the risk is itself important. In the medical field, recent literature investigates about which measures of risk are more easily understandable to also less expert people, and thus which measures might be preferable to use for presenting results on risk studies (e.g. Akl et al., 2011; Zipkin et al., 2014)
In this work we contribute to the existing literature on risk communication by analyzing data from a study on the communication of epidemiological results to the population, conducted in the high environmental risk area of Livorno, in the center of Italy. In the study, to which we refer to, from now on, as "the Livorno study", a random sample of residents in the area of Livorno was randomized to respond to different questionnaires, in which the same epidemiological results on cancer mortality were expressed by two alternative risk indexes: % excess risk (% ER) and time needed to harm (TNH) (Farinella et al., 2017). The objective of the study was to investigate whether residents' concern about the occurrence of specific health outcomes was influenced by the statistical indicator used to communicate the associated epidemiological results. Individuals' degree of concern was measured using a quantitative scale ranging from 1 to 10, where 1 means no troubled and 10 means extremely troubled. Additionally, the participants were asked to rank mortalities from three types of cancers (Sexual Glands Cancer, SGC, Thyroid Gland Cancer, TGC, Lung Cancer, LC) from the most to the least alarming one (see Farinella et al., 2017, for details on the questionnariares and the study design). Baccini et al. (2019) previously analyzed these data finding that (i) the probability of a high degree of

concern was larger under TNH than under %ER; and (*ii*) people perceived mortality from sexual glands cancer as more alarming and mortality from thyroid gland cancer as less alarming under TNH than under %ER, whereas the use of the two alternative risk indexes did not impact the rank assigned to lung cancer.
In this work we further investigate the role of the two alternative statistical indicators by analyzing the mechanisms by which they affected individuals' reactions. Specifically, we are interested in understanding whether the two alternative risk indexes affected individuals' concern degree differently depending on the answer behaviours that is, whether individuals' answers were mainly driven either by the communicated numerical/statistical information or by his/her a-priori knowledge about the the diseases and the environmental risk in the area of Livorno. In order to address this issue we take advantage of a closed-ended question introduced in the questionnaire to get information about the answer behaviour, asking the participant to auto-report the main driving factor of his/her response to the previous items: communicated numbers or a priori knowledge.
We frame our discussion within the potential outcome approach to casual inference, viewing the type of risk index as the treatment variable. From this perspective, our focus is on studying whether treatment effects on individuals' degree of concern are heterogeneous with respect to the individuals' answer behaviour. Because the auto-reported answer behaviour is a post-treatment variable, which may be affected by the treatment, we cannot conditioning on the observed value of the individuals' answer behaviour. We deal with this issue using the principal stratification framework (Frangakis and Rubin, 2002) under which the quantities of interest are principal causal effects. In our case, the principal causal effects are causal effects for latent sub-populations of units defined by joint values of individuals' answer behaviour under the two risk indexes. For inference we use a Bayesian model-based approach (Rubin, 1978).
The work is organized as follows. Section 2.2 we present the Livorno study. In Section 2.3 we re-formulate the research question of the Livorno study as a causal inference problem introducing the basic concepts of the potential outcome approach in the context of our study. In Section 2.4 we define the principal stratification of units with respect to the auto-reported answer behaviour and the principal causal effects of interest. In Section 2.5 we display Bayesian models to conduct causal inference. We then conclude presenting the results in Section 2.6 and some general observations in Section 2.7.

2.2 The Livorno study

The data come from a randomized experiment conducted on a sample of residents in the area of Livorno, an italian city located on the western coast of Tuscany. Livorno and the neighbouring municipality of Collesalvetti are classified as a high-risk environmental site according to the Seveso Directive (82/501/EC, 96/82/EC) due to the presence of a commercial harbour and several petrochemical plants producing high levels of pollution. The size of the respondents is 319. Participants are randomly assigned to reply to

different questionnaires, where two alternative risk indexes are used to communicate the same epidemiological results concerning the health profile of the Livorno-Collesalvetti population between 2001 and 2006. The risk indexes used are: the percentage of excess of risk of death in Livorno-Collesalvetti in respect to Tuscany, $\%ER = 100 * (O - E)/E$, where O is the observed number of deaths from a specific cause in the area during the period of interest, and E is the corresponding expected number of deaths, calculated according to the regional rates by age, gender and deprivation level; and the time needed to harm (TNH), i.e. the number of days one has wait on average to observe 1 death in excess in the Livorno area, taking Tuscany as the reference area, $TNH = N/(O - E)$, where N is the total number of days in the period of interest. It is worth noting that the two risk indexes provide the same information when coupled with the total number of deaths, but %ER, represents a relative measure of excess, while TNH is an absolute measure of excess.
Let Z_i denote the indicator of questionnaire/risk index assigned to person i, $i = 1, \ldots, n$, $n = 319$. For each person i the questionnaire collects information on a set of background characteristics, $\mathbf{X}_i$, including numerical skills, age, gender and smoking behaviour, and on the two following variables about the individual concern:

- The degree of concern about mortality measured on a quantitative scale ranging form 1 to 10. Here we will consider a binary version of this variable. Specifically let Y_{D_i} be the binary variable taking on value 1 if the degree of concern about mortality is larger than 5 and 0 otherwise (see Baccini et al., 2019, for a sensitivity analysis with respect to the choice of the cutoff point).

- A categorical variable, Y_{class_i}, with six categories, corresponding to the six possible alarm-based rankings $(1^{st}, 2^{nd}, 3^{rd})$ of the following three cancers: sexual glands cancer (SGC), thyroid gland cancer (TGC) and lung cancer (LC). We use the following labels: $Y_{class_i} = y$, with $y \in \mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$ where $1 = (SGC, TGC, LC)$, $2 = (SGC, LC, TGC)$, $3 = (TGC, SGC, LC)$, $4 = (TGC, LC, SGC)$, and for the last two categories $5 = (LC, SGC, TGC)$, $6 = (LC, TGC, SGC)\}$. For instance, if a participant considers SGC as the most alarming cancer and LC as the least alarming cancer, and thus ranks TGC in the second position in the alarm classification, then $(1^{st}, 2^{nd}, 3^{rd}) = (SGC, TGC, LC)$, and $Y_{class_i} = 1$.

Finally, let W_i denote a binary variable equal to 1 if participant i replies that s/he answered to the questions about the level of concern on the basis of the communicated epidemiological results and 0 if s/he replies that s/he answered on the basis of her/his a priori knowledge/experience about either diseases or pollution in the resident area. We refer to this variable as the *data-based-answer indicator*.

2.3 The Livorno study as a causal study

The aim of the study is to evaluate whether the communication of the epidemiological results using %ER-risk index versus TNH-risk index affects individuals' response

behaviour, represented by the data-based-answer indicator W, and risk perception as measured by Y_D and Y_{class}, and investigating whether the effect on Y_D and Y_{class} are heterogeneous with respect to W. We propose to reach these goals by re-formulating our research questions as a causal inference problem, where the indicator of the risk index used in the questionnaire, Z, is viewed as treatment variable and using the principal stratification framework to properly define causal effects adjusted for the post-treatment answer behaviour, W. According to the potential outcome approach (e.g. Rubin, 1974, 1978), we define potential outcomes for each post-treatment variable. Under the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980), for each unit there exist two potential outcomes for each post-treatment variables: $W_i(\%ER)$ and $W_i(TNH)$; $Y_{D_i}(\%ER)$ and $Y_{D_i}(TNH)$; and $Y_{class_i}(\%ER)$ and $Y_{class_i}(TNH)$, which are the values of W_i, Y_{D_i} and Y_{class_i} that, possibly contrary to facts, would have been observed for unit i under %ER-risk and TNH-risk, respectively.

We investigate whether the use of %ER-risk index versus TNH-risk index affects W, Y_D and Y_{class} focusing on the following causal estimands:

$$\tau_W = \underline{E}[W_i(\%\text{ER}) - W_i(\text{TNH})] = \Pr(W_i(\%\text{ER}) = 1) - \Pr(W_i(\text{TNH}) = 1), \quad (2.1)$$

$$\tau_{Y_D} = \underline{E}[Y_{D_i}(\%\text{ER}) - Y_{D_i}(\text{TNH})] = \Pr(Y_{D_i}(\%\text{ER}) = 1) - \Pr(Y_{D_i}(\text{TNH}) = 1), \quad (2.2)$$

$$\tau_{Y_{class}}(A) = \Pr(Y_{class_i}(\%\text{ER}) \in A) - \Pr(Y_{class_i}(\text{TNH}) \in A), \quad (2.3)$$

where A is a non empty subset of the power set of $\mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$. We, also, denote with $A_{SGC} := \{1, 2\}$, $A_{TGC} := \{3, 4\}$ and with $A_{LC} := \{5, 6\}$ three of the possible empty subsets of $\mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$.
For instance, the lung cancer is classified as the most alarming cause of death if $Y_{class_i} \in A_{LC}$ therefore $\tau_{Y_{class}}(A_{LC})$ is the causal effect of the %ER versus TNH risk index on the classification of lung cancer as the most alarming cause of death.
For any individual, only one of the two potential outcomes is observed, depending on the questionnaire the individual is assigned to. Let:

$$\begin{aligned} W_i^{obs} &= W_i(\%ER)\mathbb{I}\{Z_i = \%ER\} + W_i(TNH)\mathbb{I}\{Z_i = TNH\} \\ Y_{D_i}^{obs} &= Y_{D_i}(\%ER)\mathbb{I}\{Z_i = \%ER\} + Y_{D_i}(TNH)\mathbb{I}\{Z_i = TNH\} \\ Y_{class_i}^{obs} &= Y_{class_i}(\%ER)\mathbb{I}\{Z_i = \%ER\} + Y_{class_i}(TNH)\mathbb{I}\{Z_i = TNH\} \end{aligned}$$

denote the observed variables. Being the type of questionnaire/risk index randomly assigned, treatment assignment does not depend on either potential outcomes or covariates, and simple comparisons of the observed groups defined by type of risk index provide valid estimates of the causal estimands in Equations (2.1), (2.2) and (2.3). Tables 2.1 and 2.2 present some summary statistics for the sample of 319 individuals included in the Livorno study, grouped by risk index type, Z_i (Table 2.1), and data-based-answer indicator W_i^{obs} (Table 2.2). As we can see in Table 2.1 the overall mean of the observed outcome $Y_{D_i}^{obs}$ is around 92%, indicating that around 92% of the respondents show a degree of concern about mortality larger than 5. Columns 2 and 3 in Table 2.1 provide

simple moment-based estimates of the quantities in Equations (2.1), (2.2) and (2.3). Thus, the difference between the observed percentages of $Y_{D_i}^{obs}$ under %ER and TNH is negative and lower than 7%.
Moreover simple moment-based estimates of the quantities in Equations (2.3) for $A = A_{SGC}$, A_{TGC}, A_{LC} suggest that $\tau_{Y_{class}}(A_{SGC})$ is negative and higher, in absolute value, than 16%; $\tau_{Y_{class}}(A_{TGC})$ is positive and around the 13%; $\tau_{Y_{class}}(A_{LC})$ is positive and equal to 3%.

2.4 Principal stratification w.r.t. data-based-answer indicator

In order to better understand the mutual role of risk index and prior knowledge in affecting the participants' answers, it is worthwhile to investigate the heterogeneity of the effects on risk perception with respect to the response behaviour identified by the data-based-answer indicator, W. Because the indicator W may be affected by the type of risk index used in communication, we deal with this issue using the principal stratification framework (Frangakis and Rubin, 2002). The (basic) principal stratification with respect to the indicator W cross-classifies subjects into four latent groups, named principal strata, defined by the joint potential values of W, $(W_i(\%\text{ER}), W_i(\text{TNH}))$. Formally, individuals can be grouped into the following four latent groups:

- Never Empirical: $NE \equiv \{i : W_i(\%\text{ER}) = W_i(\text{TNH}) = 0\}$, participants who would never answer on the bases of the communicated data irrespective of the type of risk-index:

- Always Empirical: $AE \equiv \{i : W_i(\%\text{ER}) = W_i(\text{TNH}) = 1\}$, participants who would always answer on the bases of the communicated data irrespective of the type of risk-index

- Confident with %ER-risk index: $C_{\%\text{ER}} \equiv \{i : W_i(\%\text{ER}) = 1, W_i(\text{TNH}) = 0\}$, participants who would answer on the bases of the communicated data under the %ER-risk index, but who would not answer on the bases of the communicated data under the TNH-risk index

- Confident with TNH-risk index: $C_{\text{TNH}} \equiv \{i : W_i(\%\text{ER}) = 0, W_i(\text{TNH}) = 1\}$, participants who would answer on the bases of the communicated data under the TNH-risk index, but who would not answer on the bases of the communicated data under the %ER-risk index

Let $G_i \in PS := \{NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}\}$ denote the principal stratum membership. Principal Strata are not affected by treatment assignment, therefore, we can define causal effects conditional on the principal strata, known as principal causal effects:

$$\begin{aligned} \tau_{Y_D}^g &= \underline{E}[Y_{D_i}(\%\text{ER}) - Y_{D_i}(\text{TNH}) \mid G_i = g] \\ &= \Pr(Y_{D_i}(\%\text{ER}) = 1 \mid G_i = g) - \Pr(Y_{D_i}(\text{TNH}) = 1 \mid G_i = g), \end{aligned} \tag{2.4}$$

$$\tau^{g}_{Y_{class}}(A) = \Pr(Y_{class_i}(\%\text{ER}) \in A \mid G_i = g) - \Pr(Y_{class_i}(\text{TNH}) \in A \mid G_i = g), \quad (2.5)$$

for $g \in PS = \{NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}\}$.

It is worth noting that the causal effect on the data-based-answer indicator in Equation (2.1) can be re-written as difference between the probability of being confident with %ER-risk index and the probability of being confident with TNH-risk index: $\tau_W = \Pr(G_i = C_{\%\text{ER}}) - \Pr(G_i = C_{\text{TNH}})$.

Moreover, the causal effects on τ_{Y_D} and $\tau_{Y_{class}}$ in Equations (2.2) and Equation (2.3) can be re-written as weighted average of principal causal effects in Equations (2.4) and Equation (2.5) with weights the principal stratum proportions:

$$\tau_{Y_D} = \sum_{g \in PS} \tau^{g}_{Y_D} \Pr(G_i = g) \quad \text{and} \quad \tau_{Y_{class}}(A) = \sum_{g \in PS} \tau^{g}_{Y_{class}}(A) \Pr(G_i = g).$$

The principal stratification framework makes it clear that only for the AE group, comprising subjects who would always answer using the epidemiological results irrespective of whether they are communicated in terms of $\%ER$ or TNH, we can hope to learn something about the causal effect of the type of risk index on risk perception. For the $C_{\%\text{ER}}$ and C_{TNH} groups, causal effects on risk perception, Y_D and Y_{class}, are completely confounded with causal effects on W, and NE individuals never answer using the epidemiological results irrespective of whether they are communicated in terms of $\%ER$ or TNH.

We can reasonably expect that causal effects of risk index on risk perception, as measured by Y_D and Y_{class}, for NE are zero or negligible, so that, a type of exclusion restriction for NE holds. Nevertheless we do not impose a priori this type of exclusion restriction.

2.5 Models

We adopt a Bayesian model-based approach to inference, where the potential outcomes are viewed as random variables, some observed and some unobserved. Bayesian principal stratification requires to specify two sets of models: (1) a statistical model for the principal stratum membership (conditional on covariates); and (2) statistical models for the potential outcomes of the main endpoints conditional on principal stratum membership and covariates and a prior distribution (e.g. Imbens and Rubin, 1997). We conduct two parallel distinct principal stratification analyses, one for the degree of concern Y_D, and one for the ranking variable Y_{class}. It is worth noting that in the questionnaires the question about whether the individual answers either using the communicated epidemiological results or his/her a-priori knowledge/experience is related to Y_{class}. Nevertheless it is reasonable to believe that an individual uses the same criterion even when s/he answers the question about the degree of concern. Therefore we conduct a principal stratification analysis for both outcomes, with W as intermediate variable. We expect that the resulting classification of subjects into the four latent groups, $NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}$, will be similar, with differences mainly due to sampling

variability.

Principal Stratum Submodel

We specify conditional probit models for principal stratum membership $G_i \in PS = \{AE, NE, C_{\%ER}, C_{TNH}\}$, given the set of covariates $\mathbf{X}_i$:

$$\begin{array}{lcl} \pi_i(AE) & \equiv & Pr(G_i = AE \mid \mathbf{X}_i;\boldsymbol{\theta}) = Pr(G_i^*(AE) \leq 0 \mid \mathbf{X}_i;\boldsymbol{\theta}), \\ \pi_i(NE) & \equiv & Pr(G_i = NE \mid \mathbf{X}_i;\boldsymbol{\theta}) = Pr(G_i^*(AE) > 0, G_i^*(NE) \leq 0 \mid \mathbf{X}_i;\boldsymbol{\theta}), \\ \pi_i(C_{\text{TNH}}) & \equiv & Pr(G_i = C_{\text{TNH}} \mid \mathbf{X}_i;\boldsymbol{\theta}) = Pr(G_i^*(AE) > 0, G_i^*(NE) > 0\, G_i^*(C_{\text{TNH}}) \leq 0 \mid \mathbf{X}_i;\boldsymbol{\theta}), \end{array}$$

where

$$G_i^*(g) = \alpha_g + \mathbf{X}_i'\boldsymbol{\alpha}_g^{(X)} + \epsilon_{i,g}$$

are latent variables with $\epsilon_{i,AE}$, $\epsilon_{i,NE}$ and $\epsilon_{i,C_{\text{TNH}}}$ independent random variable with standard Normal distribution. Clearly, we have

$$\begin{array}{lcl} \pi_i(C_{\%\text{ER}}) & \equiv & Pr(G_i = C_{\%\text{ER}} \mid \mathbf{X}_i;\boldsymbol{\theta}) \\ & = & 1 - Pr(G_i = AE \mid \tilde{S}_i, \mathbf{X}_i;\boldsymbol{\theta}) - Pr(G_i = NE \mid \mathbf{X}_i;\boldsymbol{\theta}) - Pr(G_i = C_{\text{TNH}} \mid \mathbf{X}_i;\boldsymbol{\theta}) \end{array}$$

The prior distributions for the principal stratum submodel are $\boldsymbol{\alpha}_g \equiv [\alpha_g, \boldsymbol{\alpha}_g^{(X)}] \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\alpha}_g}, \underline{\sigma}^2_{\boldsymbol{\alpha}_g}\mathbf{I})$, for $g = AE, NE, C_{\text{TNH}}$, where $\underline{\mu}_{\boldsymbol{\alpha}_g}$ are vector of zeros and $\underline{\sigma}^2_{\boldsymbol{\alpha}_g}$ are hyperparameters set at 100.

Submodel for Y_D

For the submodel on Y_D, we use a probit specification, given principal stratum membership and covariates:

$$\pi^D_{i,zg} \equiv \Pr(Y_{D_i}(z) = 1 \mid G_i = g, \mathbf{X}_i;\boldsymbol{\theta}) = \Pr(Y^*_{D_i}(z) > 0 \mid G_i = g, \mathbf{X}_i;\boldsymbol{\theta})$$

where

$$Y^*_{D_i}(z) = \beta_{zg} + \mathbf{X}_i'\boldsymbol{\beta}^{(X)}_{zg} + \epsilon_{i,zg},$$

with $\epsilon_{i,zg}$ i.i.d. random variables with $\epsilon_{i,zg} \sim \mathcal{N}(0,1)$, $z = \%ER, TNH$, and $g = NE$, AE, $C_{\%\text{ER}}$, C_{TNH}. Because of the relative small sample size, in order to avoid imprecise estimates, we assume equality of the slope coefficients: $\boldsymbol{\beta}^{(X)} \equiv \boldsymbol{\beta}^{(X)}_{zg}$, for all pairs (z,g) $\in \{(\%ER, NE), (TNH, NE), (\%ER, AE), (TNH, AE), (\%ER, C_{\%\text{ER}}), (TNH, C_{\%\text{ER}}), (\%ER, C_{\text{TNH}}), (TNH, C_{\text{TNH}})\}$. We assume that the parameters of the degree of concern submodel β_{zg}, $z = \%ER, TNH$, and $g = NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}$, and $\boldsymbol{\beta}^{(X)}$ are a-priori independent and independent of $\boldsymbol{\alpha}_g$, $g = NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}$. Their prior distributions are $\beta_{zg} \sim \mathcal{N}(\underline{\mu}_{\beta_{zg}}, \underline{\sigma}^2_{\beta_{zg}})$; and $\boldsymbol{\beta}^{(X)} \sim \mathcal{N}(\underline{\mu}_{\boldsymbol{\beta}^{(X)}}, \underline{\sigma}^2_{\boldsymbol{\beta}^{(X)}}\mathbf{I})$, where $\underline{\mu}_{\beta_{zg}} = 0$ $\underline{\mu}_{\boldsymbol{\beta}^{(X)}}$ is a vector of zeros, and $\underline{\sigma}^2_{\beta_{zg}}$ and $\underline{\sigma}^2_{\boldsymbol{\beta}^{(X)}}$ are set at 100.

Submodel for Y_{class}

We specify conditional probit models for Y_{class}, given principal stratum membership and covariates,

$$\begin{aligned}
\pi_{i,zg}(1) &\equiv Pr(Y_{class_i}(z)=1 \mid G_i=g,\mathbf{X}_i;\boldsymbol{\theta}) = Pr(Y^{*,1}_{class_i}(z)\leq 0 \mid G_i=g,\mathbf{X}_i;\boldsymbol{\theta}),\\
&\quad \text{and for the other categories j=2,3,4,5}\\
\pi_{i,zg}(j) &\equiv Pr(Y_{class_i}(z)=j \mid, G_i=g,\mathbf{X}_i;\boldsymbol{\theta})\\
&= \prod_{k<j} Pr(Y^{*,k}_{class_i}(z)>0 \mid G_i=g,\mathbf{X}_i,\boldsymbol{\theta}) \times Pr(Y^{*,j}_{class_i}(z)\leq 0 \mid G_i=g,\mathbf{X}_i;\boldsymbol{\theta}),
\end{aligned}$$

where

$$Y^{*,j}_{class_i}(z) = \gamma^{(j)}_{zg} + \mathbf{X}'_i\,\boldsymbol{\gamma}^{(X,j)}_{zg} + \epsilon^{(j)}_{i,zg}, \quad \epsilon^{(j)}_{i,zg} \sim \text{ i.i.d. } \mathcal{N}(0,1),$$

for $j=1,2\ldots,5$, $z=\%ER, TNH$ and $g\in\{NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}\}$, and

$$\begin{aligned}
\pi_{i,zg}(6) &\equiv Pr(Y_{class_i}(z)=6 \mid G_i=g,\mathbf{X}_i;\boldsymbol{\theta}) = 1-\sum_{j=1}^{5} Pr(Y_{class_i}(z)=j \mid G_i=g,\mathbf{X}_i;\boldsymbol{\theta})\\
&= \prod_{j=1}^{5} Pr(Y^{*,j}_{class_i}(z)>0 \mid G_i=g,\mathbf{X}_i;\boldsymbol{\theta}).
\end{aligned}$$

We again assume equality of the slope coefficients: for $j=1,\ldots,5$, $\boldsymbol{\gamma}^{(X,j)} \equiv \boldsymbol{\gamma}^{(X,j)}_{zg}$, for all couples $(z,g)\in\{(\%ER,NE),\ (TNH,NE),\ (\%ER,AE),\ (TNH,AE),\ (\%ER,C_{\%\text{ER}}),\ (TNH,C_{\%\text{ER}}),\ (\%ER,C_{\text{TNH}}),\ (TNH,C_{\text{TNH}})\}$. We assume that the parameters of the rank submodel, γ_{zg}, $z=\%ER$, TNH, and $g=NE$, AE, $C_{\%\text{ER}}$, C_{TNH}, and $\boldsymbol{\gamma}^{(X,j)}$, $j=1,2,3,4,5$, are a-priori independent and independent of $\boldsymbol{\alpha}_g$, $g=NE, AE, C_{\%\text{ER}}, C_{\text{TNH}}$. Their prior distributions are $\gamma_{zg}\sim\mathcal{N}(\underline{\mu}_{\gamma_{zg}}, \underline{\sigma}^2_{\gamma_{zg}})$; and $\boldsymbol{\gamma}^{(X,j)}\sim\mathcal{N}(\underline{\mu}_{\boldsymbol{\gamma}^{(X,j)}}, \underline{\sigma}^2_{\boldsymbol{\gamma}^{(X,j)}}\mathbf{I})$, where $\underline{\mu}_{\gamma_{zg}}=0$ $\underline{\mu}_{\boldsymbol{\gamma}^{(X,j)}}$ are vectors of zeros, and $\underline{\sigma}^2_{\gamma_{zg}}$ and $\underline{\sigma}^2_{\boldsymbol{\gamma}^{(X,j)}}$ are set at 100.
Posterior distributions are simulated from Markov chain Monte Carlo (MCMC) algorithms, which involve Gibbs sampler steps and use the Data Augmentation (DA) method of Tanner and Wong (1987) to impute at each iteration the missing principal stratum membership. The chains are run for 10 000 iterations burning the first 2 000 iterations.

2.6 Results

We present a summary of the posterior distributions of the estimands defined in Equations (2.1), (2.2) and (2.4), (2.3) and (2.5). In Tables 2.3 and 2.4 we report the results provided by the analysis on the outcome Y_D, in Table 2.5 and in Table 2.6 the results from the analysis related to the outcome Y_{class}. For this last analysis we also report in Table 2.7 summary statistics of the posterior distributions of the estimands defined in Equations (2.3) and (2.5), for subsets $A_{l_1>l_2}$, where $l_1\neq l_2$, and $l_1,l_2\in\{SGC,TGC,LC\}$,e.g. $A_{SGC>TGC}$ is defined by the categories $y\in\mathcal{Y}=\{1\ldots6\}$ such that the rank assigned

to SGC is greater than the rank assigned to TGC. In each table, we report posterior median, 95% and 90% Posterior Credible Intervals (PCIs).

We first discuss the results from the principal stratification analysis for degree of concern, Y_D. The proportion of principal strata (Column 2 Table 2.3) show that there exists a high proportion of Never Empirical (NE) participants, suggesting that most of the participants would not have answered to the questions relying on communicated data, irrespective by the risk index type used.

Causal estimand τ_W provides information about the effect of the two alternative risk indexes on the data-based answer indicator W. Its posterior median is -0.012 (Table 2.3, Column 2), with PCI(95%)=$(-0.052, 0.002)$ and PCI(90%)=$(-0.044, -0.002)$ indicating that the proportion of people that base their response about the level of concern on the communicated data is slightly higher under TNH than under % ER, even if the PCIs include the zero. The median of the posterior distribution of τ_{Y_D} is negative, -0.057, suggesting that participants show more concern if TNH is used instead of %ER.

The causal effects are negative also within each principal stratum (see $\tau^g_{Y_D}$, with $g \in \{AE, NE, C_{\%\text{ER}}, C_{\text{TNH}}\}$), but, if we exclude the AE-stratum, the PCIs are wide and include the null hypothesis of no effect. In the AE-stratum the posterior median of the effect is the highest, around 0.08; this value can be interpreted as the net effect of TNH versus %ER.

Let consider now the results provided by the principal stratification analysis for Y_{class}. The posterior median of τ_W is greater, in absolute value, than that one obtained by the previous analysis on Y_D and the PCIs cover only negative values: it is -0.057 with PCI(95%)=$(-0.136, -0.015)$ and PCI(90%)=$(-0.123, -0.021)$. This indicates that there exists a difference in the data-based answer indicator W_i if one risk index is used instead of an other. In particular, the proportion of people that base their response on the communicated data is higher under TNH than under %ER.

Column 2 of Table 2.5 shows the posterior median of the estimands $\tau_{Y_{class}}(A_l)$ for $l \in \{SGC, TGC, LC\}$. We can observe that the posterior median of $\tau_{Y_{class}}(A_{SGC})$ is negative and that of $\tau_{Y_{class}}(A_{TGC})$ is positive, both with PCIs that do not include the zero ($\tau_{Y_{class}}(A_{TGC}) = 0.120$, $\tau_{Y_{class}}(A_{LC}) = 0.028$). The posterior median of $\tau_{Y_{class}}(A_{LC})$ is positive, but with PCI including zero.

Let focus now on the principal causal effect for the AE-stratum which we remind to be the stratum who allows us to learn about the causal effect of the risk index on the ranking assigned to the three causes of death (Table 2.6 Column 2). The posterior distribution of $\tau^{\text{AE}}_{Y_{class}}(A_{LC})$ seems to indicate that there is no effect of the indicator of the risk index, Z, on the probability of assigning rank 1 to LC. On the contrary, there is an effect of the risk index indicator on the probability of indicating SGC and TGC as the most alarming causes of death, as shown by the PCIs estimated for $\tau^{\text{AE}}_{Y_{class}}(A_{SGC})$ and $\tau^{\text{AE}}_{Y_{class}}(A_{TGC})$. In particular, the probability of assigning rank 1 to SGC is higher under TNH than under %ER. Viceversa for TGC.

Finally, let consider the subsets $A_{l_1 > l_2}$ which define the all possible comparisons between ranks of two of the three types of cancers. Table 2.7 shows summary statistics of

the posterior distributions of the causal effects $\tau_{Y_{\text{class}}}(A_{l_1>l_2})$. Specifically, first rows of Column 2 report the posterior median of the expected values under $Z = \%ER$ and $Z = TNH$, respectively. Results suggest that participants consider as more alarming LC than SGC or TGC under both types of risk index, this explains the posterior medians of $\tau_{Y_{\text{class}}}(A_{LC>SGC})$ and $\tau_{Y_{\text{class}}}(A_{LC>TGC})$ lower than the others. The $\tau_{Y_{\text{class}}}(A_{LC>SGC})$ is positive, around 0.059, with PCI including the zero, suggesting an higher probability of zero effect of the risk index used on communicating data, irrespective by which one of the two types. The $\tau_{Y_{\text{class}}}(A_{LC>TGC})$ is negative, -0.077. The PCIs in this case suggest an effect of the risk index used, on communicating data: PCI(95%)=$(-0.159, 0.004)$ with the upper limit close to zero, but with the $\sim 97\%$ of values of the posterior for $\tau_{Y_{\text{class}}}(A_{LC>TGC})$ lower than zero.
Considering the others posterior medians of the probabilities (Table 2.7), we can conclude that after LC, the classification keeps the SGC as more alarming than TGC, $\Pr(Y_{class_i}(\%\text{ER}) \in A_{SGC>TGC})$ and $\Pr(Y_{class_i}(\text{TNH}) \in A_{SGC>TGC})$ are both higher than the 50%; $\tau_{Y_{\text{class}}}(A_{SGC>TGC})$ is negative, -0.088. The PCIs also in this case suggest an effect of the risk index used, on ranking SGC as more alarming cause of death than TGC.

According the posterior distribution of the rank assigned to the three causes of death under TNH and %ER, which can be derived from the posterior distribution of Y_{class}, we obtain the corresponding Cumulatives Rankograms (CRks) (Figure 2.1). The CRk for a specific cancer l represents the probability that cancer l is classified among the k most worrisome ones, where k ranges from one to three. For each CRk, we calculate also the surface under it, so called SUCRA Salanti et al. (2011), which can also be interpreted as the average proportion of cancers classified as least alarming than cancer l. Thus, a value of SUCRA(l) equal to one means cancer l always ranks first. Table 2.8 reports posterior median and PCIs for the three SUCRAs under both treatment arms. SUCRA(LC) posterior median is around 75% under both treatment arms, suggesting that LC has high probability to be classified as the most alarming cause of deaths among the three types of cancer. Figure 2.1 shows the same conclusion graphically: the CRk of LC (black curve) is above the others under both treatment arms. Considering the CRk(SGC) (red curve) and the CRk(TGC) (green curve), they switch under %ER after rank 2, suggesting that the rank of SGC and TGC as 2^{nd} or 3^{rd} vary if one risk index is used instead an other.

2.7 Conclusions

In this work we explore the role of the numerical format in communicating epidemiological results to the population and its relationship with the a priori knowledge.
Motivated by "the Livorno study" we address our aim by re-formulating it as a problem of causal inference, using the principal stratification approach within the potential outcome framework.
The causal effects of using different risk indexes on the level of concern of the respondents,

Y_D, and on the ranking assigned to three diseases, Y_{class}, may be confounded with the causal effect on the "response behaviour" as identified by the indicator W. For this reason we define four principal strata with respect to the couple of potential outcomes of the indicator W: Never Empirical (NE), Always Empirical (AE), Confident with %ER risk index ($C_{\%ER}$) and Confident with TNH risk index (C_{TNH}) strata, which identify four potential profile of individuals as the related names well describe.
The AE-stratum is then identified as the only stratum from which we can learn about the effect of the relative effect of the two risk indicators.
In fact, individuals belonging to AE-stratum look at communicated data irrespective of the numerical format of the communicated results, therefore in this stratum the causal effects on Y_D and Y_{class} are not confounded with the causal effect on response behaviour, W.
The results show that the highest proportion of persons, $\sim 63 - 69\%$ (percentage of persons belonging to the NE-stratum) do not consider numbers in any form, around 25% (percentage of persons belonging to the AE-stratum) of them answer on the basis of the reported results. Focusing on those persons who always look at communicated data, i.e. AE persons, the two indicators, %ER and TNH risk index, induced different levels of concern and different ranking. These results raises issues which are important in the prospective of facilitating public health communication and promoting equal access to information across society.

Table 2.1: Livorno study: Means/Prevalences for the sample and classified by risk index, Z_i.

Variables ($n = 319$)	All	$Z_i = \%ER$	$Z_i = TNH$
Risk index ($Z_i = TNH$)		0.498	0.502
Data-based-answer indicator (W_i^{obs})	0.313	0.352	0.275
Background variables			
Age	51.091	51.176	51.006
Gender (Male)	0.495	0.516	0.475
Smoking behaviour			
No smoker	0.571	0.553	0.588
Ex smoker	0.223	0.220	0.225
Smoker	0.207	0.226	0.188
Numerical skills	0.790	0.767	0.812
Outcome variables			
Degree of concern (Y_D^{obs})	0.917	0.885	0.949
Rank (Y_{class})			
1 = (SGC, TGC, LC)	0.041	0.013	0.069
2 = (SGC, LC, TGC)	0.135	0.082	0.188
3 = (TGC, SGC, LC)	0.088	0.132	0.044
4 = (TGC, LC, SGC)	0.091	0.113	0.069
5 = (LC, SGC, TGC)	0.367	0.403	0.331
6 = (LC, TGC, SGC)	0.279	0.258	0.300

Table 2.2: Livorno study: Means/Prevalences for the sample grouped by risk index, Z_i, and observed data-based-answer indicator, W_i^{obs}.

Variables ($n = 319$)	$Z_i = \%ER$		$Z_i = TNH$	
	$W_i^{obs} = 0$	$W_i^{obs} = 1$	$W_i^{obs} = 0$	$W_i^{obs} = 1$
Data-based-answer indicator (W_i^{obs})	0.645	0.352	0.725	0.275
Background variables				
Age	52.427	48.875	52.526	47.000
Gender (Male)	0.505	0.536	0.491	0.432
Smoking behaviour				
No smoker	0.592	0.482	0.534	0.727
Ex smoker	0.175	0.304	0.250	0.159
Smoker	0.233	0.214	0.216	0.114
Numerical skills	0.728	0.839	0.802	0.841
Outcome variables				
Degree of concern (Y_D^{obs})	0.893	0.821	0.905	0.977
Rank (Y_{class})				
1 = (SGC, TGC, LC)	0.013	0.000	0.031	0.038
2 = (SGC, LC, TGC)	0.057	0.025	0.062	0.125
3 = (TGC, SGC, LC)	0.019	0.113	0.031	0.012
4 = (TGC, LC, SGC)	0.063	0.050	0.031	0.038
5 = (LC, SGC, TGC)	0.296	0.107	0.300	0.031
6 = (LC, TGC, SGC)	0.201	0.057	0.269	0.031

Table 2.3: Summary statistics of the posterior distributions of estimands for Y_D: Posterior Median, 95% and 90% Posterior Credible Intervals (PCIs)

ESTIMANDS	50%	PCI(95%)	PCI(90%)
$\Pr(G_i = AE)$	0.286	$(0.234, 0.338)$	$(0.243, 0.320)$
$\Pr(G_i = NE)$	0.690	$(0.631, 0.740)$	$(0.641, 0.723)$
$\Pr(G_i = C_{\%ER})$	0.003	$(0.000, 0.019)$	$(0.000, 0.013)$
$\Pr(G_i = C_{TNH})$	0.015	$(0.000, 0.066)$	$(0.001, 0.043)$
τ_W	-0.012	$(-0.052, 0.002)$	$(-0.044, -0.002)$
τ_{Y_D}	-0.057	$(-0.125, 0.001)$	$(-0.114, -0.008)$

Table 2.4: Summary statistics of the posterior distributions of principal causal effects for Y_D: Posterior Median, 95% and 90% Posterior Credible Intervals (PCIs)

ESTIMANDS	50%	C.I(95%)	C.I(90%)
$\tau_{Y_D}^{\mathrm{AE}}$	-0.079	$(-0.191, 0.000)$	$(-0.171, -0.011)$
$\tau_{Y_D}^{\mathrm{NE}}$	-0.034	$(-0.109, 0.042)$	$(-0.096, 0.015)$
$\tau_{Y_D}^{\%\mathrm{ER}}$	-0.029	$(-1.000, 0.883)$	$(-1.000, 0.717)$
$\tau_{Y_D}^{\mathrm{TNH}}$	-0.026	$(-1.000, 0.869)$	$(-1.000, 0.704)$

Table 2.5: Summary statistics of the posterior distributions of estimands for Y_{class}: Posterior Median, 95% and 90% Posterior Credible Intervals (PCIs)

ESTIMANDS	50%	PCI(95%)	PCI(90%)
$\Pr(G_i = AE)$	0.253	$(0.187, 0.310)$	$(0.197, 0.300)$
$\Pr(G_i = NE)$	0.634	$(0.560, 0.696)$	$(0.572, 0.686)$
$\Pr(G_i = C_{\%\mathrm{ER}})$	0.026	$(0.011, 0.049)$	$(0.013, 0.045)$
$\Pr(G_i = C_{\mathrm{TNH}})$	0.082	$(0.038, 0.173)$	$(0.043, 0.158)$
τ_W	-0.057	$(-0.136, -0.015)$	$(-0.123, -0.021)$
$\tau_{Y_{class}}(A_{SGC})$	-0.148	$(-0.226, -0.072)$	$(-0.212, -0.084)$
$\tau_{Y_{class}}(A_{TGC})$	0.120	$(0.046, 0.195)$	$(0.059, 0.184)$
$\tau_{Y_{class}}(A_{LC})$	0.028	$(-0.064, 0.118)$	$(-0.048, 0.102)$

Table 2.6: Summary statistics of the posterior distributions of principal causal effects related to Y_{class}: Posterior Median, 95% and 90% Posterior Credible Intervals (PCIs)

ESTIMANDS	50%	PCI(95%)	PCI(90%)
$\tau^{\text{AE}}_{Y_{class}}(A_{SGC})$	-0.627	$(-0.784, -0.442)$	$(-0.761, -0.474)$
$\tau^{\text{NE}}_{Y_{class}}(A_{SGC})$	-0.035	$(-0.136, 0.062)$	$(-0.119, 0.045)$
$\tau^{\%\text{ER}}_{Y_{class}}(A_{SGC})$	0.882	$(0.153, 1.000)$	$(0.237, 1.000)$
$\tau^{\text{TNH}}_{Y_{class}}(A_{SGC})$	0.158	$(-0.451, 0.543)$	$(-0.318, 0.481)$
$\tau^{\text{AE}}_{Y_{class}}(A_{TGC})$	0.49	$(0.250, 0.937)$	$(0.29, 0.892)$
$\tau^{\text{NE}}_{Y_{class}}(A_{TGC})$	0.091	$(0.011, 0.161)$	$(0.024, 0.15)$
$\tau^{\%\text{ER}}_{Y_{class}}(A_{TGC})$	-0.797	$(-1.000, -0.07)$	$(-1.000, -0.143)$
$\tau^{\text{TNH}}_{Y_{class}}(A_{TGC})$	-0.496	$(-0.993, -0.027)$	$(-0.948, -0.135)$
$\tau^{\text{AE}}_{Y_{\text{LC}}}$	0.127	$(-0.269, 0.348)$	$(-0.198, 0.318)$
$\tau^{\text{NE}}_{Y_{\text{LC}}}$	-0.054	$(-0.166, 0.061)$	$(-0.147, 0.042)$
$\tau^{\%\text{ER}}_{Y_{\text{LC}}}$	0.000	$(-0.745, 0.100)$	$(-0.623, 0.044)$
$\tau^{\text{TNH}}_{Y_{\text{LC}}}$	0.399	$(-0.167, 0.964)$	$(-0.084, 0.828)$

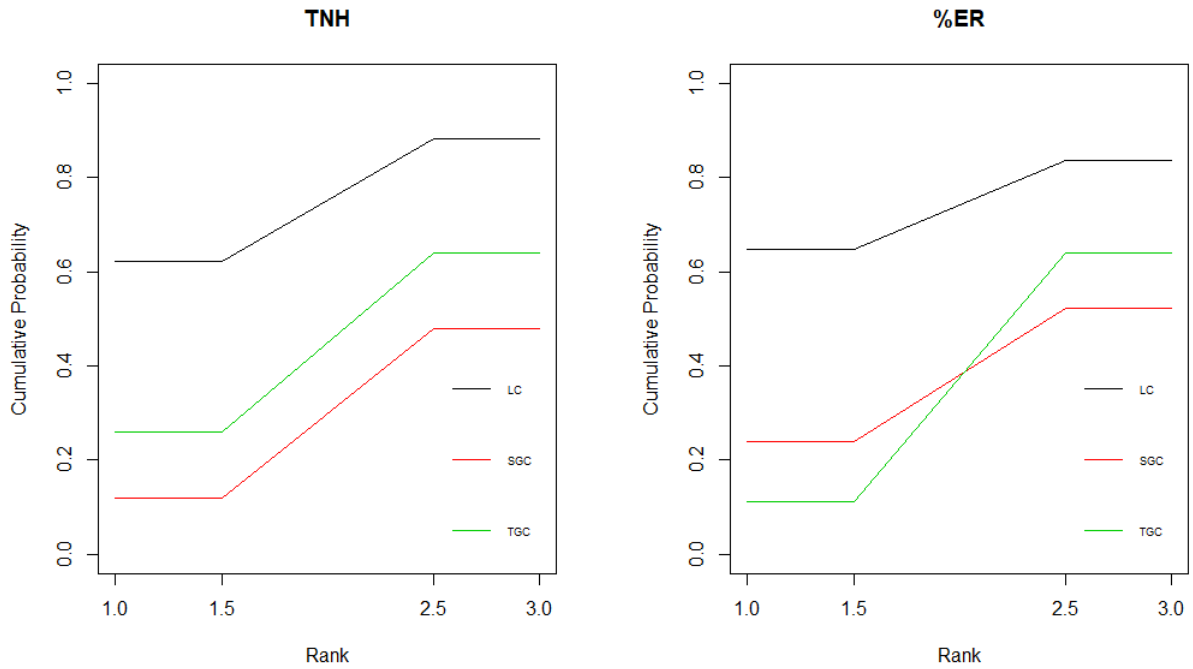
Table 2.7: Summary statistics for estimands $\Pr(Y_{class_i}(\%\text{ER}) \in A_{l1>l2})$, $\Pr(Y_{class_i}(\text{TNH}) \in A_{l1>l2})$ $\tau_{Y_{\text{class}}}(A_{l1>l2})$, with $l_1 \neq l_2$, $l_1, l_2 \in \{SGC, TGC, LC\}$: Posterior Median, 95% and 90% Posterior Credible Intervals (PCIs) .

ESTIMANDS	50%	PCI(95%)	PCI(90%)
$\Pr(Y_{class_i}(\%\text{ER}) \in A_{LC>SGC})$	0.752	$(0.686, 0.810)$	$(0.698, 0.802)$
$\Pr(Y_{class_i}(\%\text{ER}) \in A_{LC>TGC})$	0.735	$(0.673, 0.797)$	$(0.683, 0.787)$
$\Pr(Y_{class_i}(\%\text{ER}) \in A_{SGC>LC})$	0.248	$(0.19, 0.314)$	$(0.198, 0.302)$
$\Pr(Y_{class_i}(\%\text{ER}) \in A_{SGC>TGC})$	0.502	$(0.432, 0.578)$	$(0.443, 0.566)$
$\Pr(Y_{class_i}(\%\text{ER}) \in A_{TGC>SGC})$	0.498	$(0.422, 0.568)$	$(0.434, 0.557)$
$\Pr(Y_{class_i}(\%\text{ER}) \in A_{TGC>LC})$	0.265	$(0.203, 0.327)$	$(0.213, 0.317)$
$\Pr(Y_{class_i}(\text{TNH}) \in A_{LC>SGC})$	0.692	$(0.628, 0.752)$	$(0.637, 0.742)$
$\Pr(Y_{class_i}(\text{TNH}) \in A_{LC>TGC})$	0.814	$(0.752, 0.868)$	$(0.763, 0.859)$
$\Pr(Y_{class_i}(\text{TNH}) \in A_{SGC>LC})$	0.308	$(0.248, 0.372)$	$(0.258, 0.363)$
$\Pr(Y_{class_i}(\text{TNH}) \in A_{SGC>TGC})$	0.591	$(0.516, 0.663)$	$(0.527, 0.652)$
$\Pr(Y_{class_i}(\text{TNH}) \in A_{TGC>LC})$	0.409	$(0.337, 0.484)$	$(0.348, 0.473)$
$\Pr(Y_{class_i}(\text{TNH}) \in A_{TGC>SGC})$	0.186	$(0.132, 0.248)$	$(0.141, 0.237)$
$\tau_{Y_{\text{class}}}(A_{LC>SGC})$	0.059	(-0.025, 0.143)	(-0.011, 0.131)
$\tau_{Y_{\text{class}}}(A_{LC>TGC})$	-0.077	(-0.159, 0.004)	(-0.145, -0.01)
$\tau_{Y_{\text{class}}}(A_{SGC>LC})$	-0.059	(-0.143, 0.025)	(-0.131, 0.011)
$\tau_{Y_{\text{class}}}(A_{SGC>TGC})$	-0.088	(-0.190, 0.020)	(-0.176, 0.002)
$\tau_{Y_{\text{class}}}(A_{TGC>SGC})$	0.088	(-0.020, 0.190)	(-0.002, 0.176)
$\tau_{Y_{\text{class}}}(A_{TGC>LC})$	0.077	(-0.004, 0.159)	(0.011, 0.145)

Table 2.8: Posterior median, PCI(95%) and PCI(90%) of SUCRA, under TNH and %ER, for the three types of cancer.

	TNH			%ER		
	50%	PCI(95%)	PCI(90%)	50%	PCI(95%)	PCI(90%)
SUCRA(LC)	0.752	(0.702 0.799)	(0.710 0.791)	0.743	(0.688 0.794)	(0.697 0.787)
SUCRA(SGC)	0.449	(0.394 0.506)	(0.403 0.496)	0.376	(0.330 0.422)	(0.338 0.416)
SUCRA(TGC)	0.298	(0.248 0.353)	(0.256 0.343)	0.381	(0.322 0.440)	(0.331 0.430)

Figure 2.1: Cumulative rankograms for Lung cancer (black curve) Sexual Glands cancer (red curve), Thyroid gland cancer (green curve) mortalities, under TNH (on the left) and %ER (on the right).



Chapter 3

Stratification via Propensity Score and Bayesian Additive Regression Trees approaches to estimate PATT from RCT results

Key words: Propensity Score, BART, Randomization, Causal Effect, External Validity, Potential outcome, Estimator.

3.1 Introduction

Randomized Controlled Trials (RCTs) have a lot of benefits, e.g. they ensure the absence of manipulation of the assignment of the treatment removing any selection bias from the sample. On the other hand the results provided by RCTs may not have external validity for the population of interest (Target Population, TP).
Several factors could be related to this lack: gap between the RCT sample and the TP due the eligibility criteria in the trial, i.e. unbalancing of the observed covariates between RCT sample and TP; difference of the treatment protocols and interventions, or different level of treatment between RCT sample and TP.
In case for which results obtained by RCT are generalizable to the TP, how the Population Average Treatment effect on Treated (PATT) can be estimated from RCT?
Existing approaches to estimate PATT require a sub-classification or matching process of the treated and controls individuals within the RCT.
Define subgroups, such that within each one the characteristics of the membership are similar, needs to established criteria which may provide bias estimates of the causal effect of interest. Approaches like the pair-matching (1-1, in each subclass one treated is matched with one control), matching with multiple controls (1-n_c, in each subclass one treated is matched to n_c controls) or the full matching ($n_t - 1$ or $1 - n_c$, in each subclass n_t treated are matched to one control or one treated is matched to n_c controls) [Rosenbaum (2019)] have been employed in study related to the generalizability of results from RCT to a TP [Stuart et al. (2001), Stuart (2010), Cole and Stuart (2010)].
Hartman et al. (2015) exploit a new matching approach, the genetic matching (a generalization of previous matching approaches, which uses an evolutionary search algorithm developed by Diamond and Sekhon (2013) to maximize the balance of observed covariates

across matched treated and controls individuals). This last one approach has been shown in Diamond and Sekhon (2013) to perform better, in terms of bias reductions of the estimates, than other Machine Learning Algorithms as random forests and boosted classification and regression trees (CART). Nevertheless, it requires high computational costs, dependent by the sample size, which may not be feasible for large samples.
Other approaches proposed to estimate the PATT involve the Bayesian Additive Regression Trees (BART) algorithm introduced by Chipman et al. (2010) (Nethery et al., 2019, e.g.). BART approach is different by the previous ones; it does not require any sub-classification procedure, it consists on using the response surface model to estimate the covariate-outcome relationships in the RCT and then it uses these estimates to predict population treatment effects in the TP. BART approach implicitly assumes: 1) the covariates overlap, Nethery et al. (2019), between the two groups, RCT sample and the TP, 2)a correct model specification used to describe associations between the outcome and the covariates. Therefore, in the hypothetical situation for which the associations between outcome and covariates differs between the two groups (RCT and TP) and/ or 1), 2) are violated, BART may provide bias estimates.
In the current work we contribute to the existing literature pointing out the exact circumstances under which the RCT results are generalizable to the TP. We show a series of complications which prevent a direct extension of the results of a RCT to the TP. Thus, we introduce two sets of estimators in order to evaluate the external validity of RCT results to the TP and to estimate the PATT from RCT.
The two sets of estimators, we propose, are based on two approaches: the Stratification via Propensity Score (SvPS) approach and BART approach. We refer to those two sets of estimators as the SvPS estimators and the BART estimators, respectively.
Specifically, for the first set of estimators, the SvPS estimators, we adapt the SvPS approach, introduced by Lunceford and Davidian (2004) to estimate the treatment effects from observational data.
Randomization is the key point of SvPS approach. The basic idea behind the SvPS approach is to construct subgroups within which the sample selection can be considered at random. The subgroups are defined by the quantiles of the Propensity Score (PS), where the PS is referred to the probability to be selected in the RCT sample conditioning on a set of covariates, unless the usual definition of it[1], Rosenbaum and Rubin (1983).
Randomization allows comparisons between the two groups defined by the indicator of membership in RCT, such that significant statistical difference of the outcome of interest between the two groups suggests the absence of exchangeability between them that is results obtained by the RCT are not generalizable.
The work is organized as follows: in Section 3.2 we introduce the notations adopted and we report the Generalizability Assumptions (Hartman et al. (2015)) required to identify the PATT from the RCT.
In Section 3.3 we illustrate the SvPS approach and BART approach; we define the two

[1]In Rosenbaum and Rubin (1983), the propensity score is defined as the probability to be assigned to the treatment conditional on a set of observed covariates.

sets of estimators, SvPS estimators and BART estimators, which allow 1) to evaluate the external validity of the results obtained from RCT (Section 3.3.1) and 2) to estimate the PATT from RCT data (Section 3.3.2).
In Section 3.4 we provide data generation and analyses to illustrate the performances of the SvPS and BART estimators described in Section 3.3; we investigate about the exact circumstances under which checking the external validity of RCT results and estimating PATT might prove difficult depend on details of the setting, e.g. misspecification of the PS model. In Section 3.5 we conduct analysis for Medicare simulated data, to assess whether results from a Clinical trial, on the effectiveness of Trans-catheter Aortic Valve Replacement (TAVR) versus the Surgical Aortic Valve Replacement (SAVR), can be appropriately adjusted and generalized to the broader Medicare population (MP). Also for this last analysis we apply the SvPS approach and the BART approach. We then conclude in Section 3.6.

3.2 Notations and Assumptions

In this Section we introduce the notations we use and the assumptions required to identify the PATT from RCT results. We denote with uppercase letters the random variables and with lowercase letters the related realizations, or constants.
Let S_i be the indicator of membership in RCT, assuming value 1 if individual i belongs to the RCT sample, 0 if s/he is a member of the TP. Let Z_i be the indicator of the treatment assigned and received (we assume compliance) which takes value 1 if individual i is assigned to the treatment and 0 otherwise. Let $\mathbf{X}_i$ be the vector of observed covariates for individual i. Finally, we denote with Y_i the post treatment variable for individual i, our primary outcome of interest.
Given the indicator S_i of the sample membership and the indicator Z_i of the treatment assigned, we denote with $B_i = (S_i, Z_i)$ the two-dimensional vector taking values $b \in \mathcal{B} := \{(0,0), (0,1), (1,0), (1,1)\}$. Given B_i, four are the potential outcomes for the post treatment variable Y_i, one for each value of B_i. Formally, given individual i, we denote these potential outcomes with $Y_i(b) = Y_i(s,z)$, for $b \in \mathcal{B}$. Specifically, $Y_i(s,z)$ is the outcome we would observe for individual i if her/his indicator membership were s and s/he received the treatment z. Comparisons between potential outcomes are causal effects. Based on the indicator S_i and Z_i we define the following casual effects:

$$\begin{aligned}
\tau_{SATE} &= \mathbb{E}[Y_i(1,1) - Y_i(1,0) \mid S_i = 1] \\
\tau_{SATC} &= \mathbb{E}[Y_i(1,1) - Y_i(1,0) \mid S_i = 1, Z_i = 0] \\
\tau_{SATT} &= \mathbb{E}[Y_i(1,1) - Y_i(1,0) \mid S_i = 1, Z_i = 1]
\end{aligned} \tag{3.1}$$

that are the Sample Average Treatment Effect (SATE), the Sample Average Treatment effect on Controlled individuals (SATC) and the Sample Average Treatment effect on Treated individuals (SATT), respectively. In a similar way we define other three casual

effects, for those belonging in the TP, as follows:

$$\begin{aligned}\tau_{PATE} &= \mathbb{E}[Y_i(0,1) - Y_i(0,0) \mid S_i = 0] \\ \tau_{PATC} &= \mathbb{E}[Y_i(0,1) - Y_i(0,0) \mid S_i = 0, Z_i = 0] \\ \tau_{PATT} &= \mathbb{E}[Y_i(0,1) - Y_i(0,0) \mid S_i = 0, Z_i = 1] \end{aligned} \tag{3.2}$$

Our main focus is on estimand τ_{PATT}. In general, in the TP the treatment is not randomly assigned, thus, the three populations causal effect estimates may be affected by bias. Hartman et. al provide assumptions under which population treatment effects can be identified from RCT data. We can also refer to this set of assumptions as the Generalizability Assumptions (G.Ass.). The assumption we start with, we consider necessary but not sufficient for the identification of the PATT from RCT data, is related to the sample mechanism.

Assumption 3.1 (**Probabilistic sample mechanism**).

$$0 < Pr(S_i = 1 \mid \mathbf{X}_i) < 1, \forall i$$

Assumption 3.1 ensures that the subgroup of treated and control individuals in the RCT sample is well representative of the TP, given a set of covariates $\mathbf{X}_i$. We assume here, for simplicity, that the set of covariates involved in the sample mechanism is the same for treated and controlled individuals.

Assumption 3.2 (**Consistency under Parallel Studies**).

$$Y_i(0,1) = Y_i(1,1) \text{ Ass.3.2.1} \tag{3.3}$$

$$Y_i(0,0) = Y_i(1,0) \text{ Ass.3.2.2} \tag{3.4}$$

Assumption 3.2 can be considered as a revisiting of the Sharp null hypothesis, Fisher (1935). For each individual i, fixing a treatment arm, z, the related potential outcome would be the same if that individual belonged to the s' group unless to the s group, $Y_i(s,z) = Y_i(s',z)$. Assumption 3.2 is also similar to Assumption of consistency under parallel design in Imai et al. (2013).
If Assumption 3.2 holds, we can write $Y_i(s,z)$ as $Y_i(z)$ for any s and z, since that the potential outcome for treatment z is independent by the indicator of membership in RCT.

Assumption 3.3 (**Strong Ignorability of Sample Assignment for Treated under treatment**).

$$(Y_i(0,1), Y_i(1,1)) \perp S_i \mid (\mathbf{X}_i, Z_i = 1)$$

An implication of Assumption 3.3 is that the expected values of potential outcomes for treated is equal to the expected value of the potential outcomes for treated for those

individuals treated in the RCT sample, adjusted by the distribution of the covariates of treated in TP. Formally we have:

$$\begin{aligned}\mathbb{E}[Y_i(s,1) \mid S_i = 0, Z_i = 1] =& \mathbb{E}_{01}\{\mathbb{E}[Y_i(s,1) \mid \mathbf{X}_i, S_i = 1, Z_i = 1]\} \\ &= \int_{\Omega_{\mathbf{X}_{01}}} \mathbb{E}[Y_i(s,1) \mid \mathbf{X}_i, S_i = 1, Z_i = 1] \cdot f(\mathbf{X}_i)\, d\mathbf{X}_i, \quad (3.5)\end{aligned}$$

for s = 0, 1. Where $\Omega_{\mathbf{X}_{01}}$ is the space of the all possible values of the vector of covariates $\mathbf{X}_i$ in the treated group of the TP; $f(\mathbf{X}_i)$ is the related density function.

Assumption 3.4 (**Strong Ignorability of Sample Assignment for Controls under treatment**).

$$(Y_i(0,0), Y_i(1,0)) \perp S_i \mid (\mathbf{X}_i, Z_i = 1)$$

Assumption 3.4 is the equivalent of Assumption 3.3 for potential outcomes for controls, for those individuals treated. In a similar way this last assumption implies that the expected values of potential outcomes for controls is equal to the expected value of the potential outcomes for controls, for individuals treated in the RCT sample adjusted by the distribution of the covariates of treated in TP. Formally we have:

$$\begin{aligned}\mathbb{E}[Y_i(s,0) \mid S_i = 0, Z_i = 1] =& \mathbb{E}_{01}\{\mathbb{E}[Y_i(s,0) \mid \mathbf{X}_i, S_i = 1, Z - i = 0]\} \\ &= \int_{\Omega_{\mathbf{X}_{01}}} \mathbb{E}[Y_i(s,0) \mid \mathbf{X}_i, S_i = 1, Z_i = 0] \cdot f(\mathbf{X}_i)\, d\mathbf{X}_i, \quad (3.6)\end{aligned}$$

for s = 0, 1. Finally, the last, but not least, assumption we need to specify in order to conduct causal inference is the Stable individual Treatment Value Assumption (SUTVA). SUTVA consists of two parts. 1) No Interference-part which states that the potential outcome of individual i, $Y_i(s,z)$, is not affected by the potential outcomes of any other individual $Y_j(s',z')$. 2) No Hidden Variations of Treatments part: no different form of a specific treatment level for a generic individual i, regardless of her/his indicator membership S_i. Under Assumptions 3.1-3.4 and SUTVA assumptions, as defined in Hartman et al. (2015), it follows the Theorem 3.1

Theorem 3.1 (**Generalizability of the results of RCT**). *If Assumption 3.2 and SUTVA hold and if*

$$\begin{aligned}&\mathbb{E}_{01}\{\mathbb{E}[Y_i(s,1) \mid \mathbf{X}_i, S_i = 0, Z_i = 1]\} - \mathbb{E}_{01}\{\mathbb{E}[Y_i(s,0) \mid \mathbf{X}_i, S_i = 0, Z_i = 1]\} \\ &= \mathbb{E}_{01}\{\mathbb{E}[Y_i(s,1) \mid \mathbf{X}_i, S_i = 1, Z_i = 1]\} - \mathbb{E}_{01}\{\mathbb{E}[Y_i(s,0) \mid \mathbf{X}_i, S_i = 1, Z_i = 1]\} \quad (3.7)\end{aligned}$$

or if Assumption 3.1, 3.3 and 3.4 hold, then $Y_i(s,z) = Y_i(s',z) = Y_i(z)$ *and it follows*

$$\tau_{PATT} = \mathbb{E}_{01}\{\mathbb{E}[Y_i(1) \mid \mathbf{X}, S_i = 1]\} - \mathbb{E}_{01}\{\mathbb{E}[Y_i(0) \mid \mathbf{X}_i, S_i = 1]\} = \mathbb{E}_{01}[\tau_{SATE_{\mathbf{X}_i}} \mid \mathbf{X}_i] \quad (3.8)$$

Theorem 3.1 shows how to identify the PATT from RCT data, that is τ_{PATT} can be derived adjusting the SATE by the covariates distribution of treated individuals in TP. Specifically, on the right hand of Equation 3.8, expression $\mathbb{E}_{01}\{\tau_{SATE_{\mathbf{X}_i}} \mid \mathbf{X}_i\}$ denotes the weighted average of $\tau_{SATE_{\mathbf{X}_i}}$ quantities with weights according to the distribution of $\mathbf{X}_i$ in the treated group of the TP. Each $\tau_{SATE_{\mathbf{X}_i}}$ refers to the average treatment effect within the subgroup of treated individuals in the RCT whose characteristics match those of the treated subgroup in the TP with values of the covariates equal to $\mathbf{X}_i$.

3.3 SvPS approach and BART approach

In the introduction of this work, we shortly described some of the existing approaches to estimate the PATT from RCT. In this section we focus on SvPS approach and BART approach.

SvPS approach, as the name suggests, basically requires a sub-classification process of the individuals in RCT and those ones in the TP and exploits the all informations provided by the RCT group.

Usually, the Propensity Score (PS) is defined as the probability to be exposed to a treatment conditioning on a set of covariates, $\mathbf{X}_i$, Rosenbaum and Rubin (1983). Here, and for the rest of this work, we refer to the PS as the probability to be selected in the RCT conditioning on a set of covariates. Formally, we denote this probability with $p_s(\mathbf{X}_i) = Pr(S_i = 1 \mid \mathbf{X}_i)$ for a general individual i.

The SvPS approach, we propose, consists on defining K strata, Q_j, for $j = 1 \dots K$, based on the quantiles q_j of the PS.

The sample mechanism is usually under the control of the researcher (in designing the study, choosing the set of covariates for which Assumption 3.1 holds) as well as under control of participants (once they are selected, they can decide to participate or not to the trial). Therefore, in most of the case the true PS is known. In case for which we do not have information about the sample mechanism, we first estimate the PS from the observed data, by assuming that the $p_s(\mathbf{X}_i)$ follows a logistic regression model $p_s(\mathbf{X}_i) = Pr(S_i = 1 \mid \mathbf{X}_i) = \dfrac{exp(\mathbf{X}_i\boldsymbol{\beta})}{(1 + exp(\mathbf{X}_i\boldsymbol{\beta}))}$. We denote with $\hat{p}_s(\mathbf{X}_i)$ the estimated PS for individual i and with $\hat{q}_j$ the estimated quantile, for $j = 1 \dots K$.

The estimated quantiles $\hat{q}_j$, $j = 1 \dots K$, are obtained sorting the individuals by their estimated PS in increasing order; the ordered group is then partitioned in K strata, $\hat{Q}_j$, such that the cardinality of each one is proportional to $\frac{N}{K}$, where N is the size of the group we estimated the PS on.

Given a stratum $\hat{Q}_j$ all individuals membership have an estimated $\hat{p}_s(\mathbf{X}_i) \leq \hat{q}_j$, where $\hat{q}_j$ is the estimated j-quantile. Within each stratum individuals membership have *similar* probabilities to be selected in the sample conditioning on a given set of covariates, that is, within each $\hat{Q}_j$, we can assume the assignment to the sample at random. Randomization allows for a direct comparison between the two groups defined by the indicator S, within each stratum.

SvPS approach requires also to fix the number of strata we want to split out individuals. Strata may contain residual confounding due to the fact that each one has individuals with *similar* propensity score, but not exactly the same.
Here, we fix the number of strata, K, at 5 as suggested in Rosenbaum and Rubin (1983)-Rosenbaum and Rubin (1984) and adopted in most published applications. This choice may avoid some residual confounding within strata.
Nevertheless, SvPS approach assumes that i) the PS model is correctly specified, ii) within each stratum each covariate is well balanced between the two groups defined by the S indicator. Checking ii) is recommended [Rosenbaum and Rubin (1984),Perkins et al. (2000)] and in case for which the balance is not achieved it is necessary to redefine the model of the PS.
BART approach is a model based approach and it consists on using a response surface model to estimate the relationship between covariates and the primary outcome in the RCT (for treated and controls separately), then it exploits the estimates to predict the outcome for those individuals in the TP according with their covariates distribution. Since that we aim to estimate the PATT, we focus on the treated in the TP and we refer to $Y_i^{fit}(s,z)_{|S_i=0,Z_i=1}$ as the general predicted potential outcome for an individual i treated in the TP. Specifically, we consider two of the four potential outcomes for an individual i treated in the TP and we refer to the related predicted value as following:

- $Y_i^{fit}(1,1)_{|S_i=0,Z_i=1}$:= the predicted potential outcome for treated in RCT for an individual i treated in TP. That is the predicted outcome individual i, membership of the TP, would have had if s/he has been sampled in the RCT and assigned to the treatment arm.
  $Y_i^{fit}(1,1)_{|S_i=0,Z_i=1}$ is obtained by 1) using a response surface model to estimate the relationship between covariates and the primary outcome in the **treated group** of RCT 2) using the estimates to fit the potential outcome for treated in RCT for those individuals treated in the TP according with their covariates distribution.

- $Y_i^{fit}(0,0)_{|S_i=0,Z_i=1}$:= the predicted potential outcome for control in TP for an individual i treated in TP. That is the predicted outcome individual i, membership of the TP, would have had if s/he has been assigned to the control arm in the TP.
  $Y_i^{fit}(0,0)_{|S_i=0,Z_i=1}$ is obtained by 1) using a response surface model to estimate the relationship between covariates and the primary outcome in the **control group** of RCT 2) using the estimates to fit the potential outcome for control in TP for those individuals treated in the TP according with their covariates distribution.

The main assumptions required in BART approach are 1) the covariates overlap between the two groups, RCT sample and the TP; 2) a correct model specification used to describe associations between the outcome and the covariates.
Next sections show in details how to apply SvPS approach and BART approach to 1) verify the external validity of RCT results 2) estimate the PATT from RCT results.

3.3.1 SvPS approach and BART approach to verify the external validity of RCT results

The external validity of RCT results suggests that individuals in RCT are exchangeable with those one in the TP. Exchangeability of the individuals is reflected in the equality of the potential outcomes distribution for the indicator of membership in the RCT, S_i, given $Z_i = z$. Nevertheless, for each individual i we almost observe one of the four potential outcomes, the other are missing. We refer to $Y_i^{obs} = \sum_{b\in\mathcal{B}} Y_i(b)\mathbb{I}(B_i = b)$ as the potential outcome observed for individual i.
Thus, we are not able to test the equality of the potential outcomes' distribution for the indicator S_i, given Z_i. In order to better judge the plausibility of the G.ass., with consequent external validity of RCT results, we define the following causal effect:

$$\tau_{(S,1)} = \mathbb{E}_{\mathbf{X}_i}\{\mathbb{E}[Y_i(1,1) - Y_i(0,1) \mid Z_i = 1, \mathbf{X}_i]\}, \tag{3.9}$$

that is the weighted mean of the expected value of the difference between the two potential outcomes $Y_i(1,1)$ and $Y_i(0,1)$. The potential outcomes in Equation 3.9 are potential outcomes for the indicator membership in RCT for those individuals treated. The weights are defined in accordance with the distribution of X_i. If Assumptions 3.1, 3.2.1, 3.3 and SUTVA hold, then it follows that the causal effect $\tau_{(S,1)}$ is zero. On the other hand, if the causal effect $\tau_{(S,1)}$ is not zero and it is also not negligible we can conclude that one or more of the assumptions 3.2.1, 3.3 and SUTVA are not plausible. Here, we focus on checking the negligible effect of $\tau_{(S,1)}$.
We first adopt SvPS approach for checking the negligible effect of $\tau_{(S,1)}$. Let G_T be the group defined by treated belonging to the RCT sample and those belonging to the TP, formally $G_T = \{i : (S_i, Z_i) = (0,1) \text{ or } (1,1)\}$, we denote with N_{G_T} the related size. Let $\mathbf{X^T}$ be the $N_{G_T} \times n_{cov}$ matrix such that the l^{th} column of the i^{th} row of $\mathbf{X^T}$ contains the value of the observed covariate l associated to individual i.
We estimate the PS on these n_{cov} covariates, assuming there are no other unmeasured covariates involved on the sample selection.
Specifically, we introduce the following estimator of the quantity $\tau_{(S,1)}$:

$$\hat{\tau}^{\text{SvPS}}_{(S,1)} = \sum_{j=1}^{K}\left(\frac{n_j}{N_{G_T}}\right)\{n_{1j}^{-1}\sum_{i=1}^{N_{G_T}} S_i Y_i^{obs}\mathbb{I}(\hat{p}_s(\mathbf{X}_i)\in\hat{Q}_j) - n_{0j}^{-1}\sum_{i=1}^{N_{G_T}}(1-S_i)Y_i^{obs}\mathbb{I}(\hat{p}_s(\mathbf{X}_i)\in\hat{Q}_j)\} \tag{3.10}$$

where n_{1j} is the number of treated individuals of the RCT sample and n_{0j} is the number of treated individuals of the TP both in stratum $\hat{Q}_j$; $\mathbb{I}(\cdot)$ is the indicator function. Quantities $\frac{n_j}{N_{G_T}}$ are proportional to $\frac{1}{K}$, therefore we can replace them with $\frac{1}{K}$ in Equation 3.10, this provides a weighted average of the causal effects $\tau_{(S,1)|\hat{Q}_j}$, which are the equivalent of $\tau_{(S,1)}$ in Equation 3.9 conditioning on stratum j.
Given the estimator $\hat{\tau}^{\text{SvPS}}_{(S,1)}$ we calculate the Confidence Interval with level confidence $(1-\alpha)\%$, $C.I._{(1-\alpha)\%}(\tau_{(S,1)})$, approximating the overall variance of $\hat{\tau}^{\text{SvPS}}_{(S,1)}$, $\sigma^2_{\hat{\tau}^{\text{SvPS}}_{(S,1)}}$, by treating

$\hat{\tau}^{\text{SvPS}}_{(S,1)}$ as the average of K independent, within stratum, sampling effect estimates as:

$$\hat{\sigma}^2_{\hat{\tau}^{\text{SvPS}}_{(S,1)}} = \frac{1}{K^2}\sum_{j=1}^{K}\hat{\sigma}^2_j$$

assuming that each stratum has equal number of individuals, where $\hat{\sigma}^2_j$ denotes the estimate of the variance of the difference between the mean of the outcomes for treated individuals in the RCT sample and those treated in the TP, in stratum j.
Formally, $\hat{\sigma}^2_j = n^{-1}_{1j}s^2_{1j} + n^{-1}_{0j}s^2_{0j}$, with $s^2_{1j} = n^{-1}_{1j}\sum_{i=1}^{N_{G_T}}\mathbb{I}(\hat{p}_{s_i} \in \hat{Q}_j)[S_i(Y^{obs}_i - \bar{y}_{1j})^2]$ $s^2_{0j} = n^{-1}_{0j}\sum_{i=1}^{N_{G_T}}\mathbb{I}(\hat{p}_{s_i} \in \hat{Q}_j)[(1-S_i)(Y^{obs}_i - \bar{y}_{0j})^2]$; $\bar{y}_{1j} = n^{-1}_{1j}\sum_{i=1}^{N_{G_T}}\mathbb{I}(i \in \hat{Q}_j)S_iY_i$; $\bar{y}_{0j} = n^{-1}_{0j}\sum_{i=1}^{N_{G_T}}\mathbb{I}(i \in \hat{Q}_j)(1-S_i)Y_i$. The estimated $C.I._{(1-\alpha\%)}(\tau_{(S,1)})$ provides an interval estimation of $\tau_{(S,1)}$, if it covers zero and its width satisfies the criteria established a priori by the research, we conclude that $\tau_{(S,1)}$ is negligible, that is Assumption 3.3 holds, otherwise we can conclude that one or more Assumptions: 3.2.1, 3.3, SUTVA, are violated.

Given the estimand $\tau_{(S,1)}$ defined in (3.9), we propose the following estimator using BART approach:

$$\hat{\tau}^{\text{BART}}_{(\text{S},1)} := \frac{1}{N_{01}}\sum_{i}^{N_{01}}Y^{fit}_i(1,1)_{|S_i=0,Z_i=1} - \frac{1}{N_{01}}\sum_{i}^{N_{01}}Z_i(1-S_i)Y^{obs}_i, \tag{3.11}$$

where N_{01} is the number of individuals treated in the TP.
The predicted value $Y^{fit}_i(1,1)_{|S_i=0,Z_i=1}$ is fitted under the Assumption 3.3, if the estimated value of $\hat{\tau}^{\text{BART}}_{(\text{S},1)}$ is not negligible this suggests that one or more of the Assumptions 3.2.1, 3.3, SUTVA, are violated.
Finally, an other assumption required to identify the PATT from RCT data is the Assumption 3.4, which involves potential outcomes for control for those treated in TP, $Y_i(0,0)_{|Z_i=1}$. This last quantity is not observed, it is missing since that for those treated in TP we can observed at most the potential outcome under treatment. Moreover, we could have in some case study the absence of the control group in the TP, Stuart et al. (2001). It follows that we cannot apply the illustrated approach above in a similar way to check if there exists a negligible effect of the indicator S for controls, for those individuals treated:

$$\tau_{(S,0)} = \mathbb{E}_{\mathbf{X}_i}\{\mathbb{E}[Y_i(0,0) - Y_i(1,0) \mid Z_i = 1, \mathbf{X}_i]\}. \tag{3.12}$$

In such situation we consider plausible Assumption 3.4, because of the randomization of the treatment in the RCT. Specifically, the randomization of the treatment in RCT guarantees the absence of manipulation of the assignment of the treatment, individuals in the treated group are exchangeable with those in the control group in RCT; potential outcomes are independent by the indicator of the treatment assigned. Given the negligible effect of the sample indicator S for those individuals treated, for transitivity we can, therefore, assume that $\tau_{(S,0)}$ is negligible. Thus, the average potential outcomes for

control for those treated in the TP, $Y_i(0,0)_{|Z_i=1}$, is obtained by the average outcomes for control for those controls in the RCT sample adjusting for the covariates distribution of the treated group in TP.

3.3.2 SvPS approach and BART approach for estimating PATT

Under the validity of Assumptions 3.1, 3.2.2, 3.3, 3.4 and SUTVA, Theorem 3.1 ensure the correct identification of PATT from RCT sample.
In this section we introduce the SvPS approach and BART approach for estimating the PATT. We first start showing the procedure using SvPS approach. Let consider the group given by the union of the treated group in the TP and the RCT sample: $G = \{i : (S_i, Z_i) = (0,1), (1,1), \text{ or } (1,0)\}$, we denote with N_G the related size. Let $\mathbf{X}^G$ the $N_G \times n_{cov}$ matrix of observed covariates for the N_G individuals in G. We estimate the PS, in case for which it is not known, including these covariates in the model, assuming there are no other unmeasured covariates involved on the sample selection.
Based on the estimated PS, we then create K strata, $\hat{Q}_j$, given by the estimated quantiles $\hat{q}_j$, for $j = 1 \dots K$.
Within each stratum the sample selection is at random, as well as the assignment mechanism to the treatment, for those in the RCT, because of the randomization of the assignment in the RCT sample. Randomization of the assignment to the treatment allows for direct comparisons between the group of treated and the group of controls for those individuals in the RCT within each quantiles $\hat{Q}_j$. The estimator for PATT using the strata created is then given by the following:

$$\hat{\tau}^{\text{SvPS}}_{\text{PATT}} = \sum_{j=1}^{K} \frac{n_{j,01}}{N_{01}} \hat{\tau}_{SATE_{|\hat{Q}_j}}, \tag{3.13}$$

where $\hat{\tau}_{SATE_{|\hat{Q}_j}}$ is the estimated SATE within stratum $\hat{Q}_j$. Formally,

$$\hat{\tau}_{SATE_{|\hat{Q}_j}} = \frac{1}{n_{j,11}} \sum_{i=1}^{n_j} Z_i S_i Y_i^{obs} \mathbb{I}(\hat{p}_s(\mathbf{X}_i) \in Q_j) - \frac{1}{n_{j,10}} \sum_{i=1}^{n_j} (1 - Z_i) S_i Y_i^{obs} \mathbb{I}(\hat{p}_s(\mathbf{X}_i) \in Q_j),$$

with $n_{j,11}$ the number of treated individuals from the RCT sample in quantile $\hat{Q}_j$; $n_{j,10}$ the number of control individuals from the RCT sample in quantile $\hat{Q}_j$; $n_j \sim \frac{N_G}{K}$ the size of quantile $\hat{Q}_j$; $n_{j,01}$ the number of treated individuals from TP in quantile $\hat{Q}_j$ and N_{01} the total number of treated individuals in the TP.
The estimator proposed applying BART approach exploits partially the information derived from RCT. It is defined as follows:

$$\hat{\tau}^{\text{BART}}_{\text{PATT}} := \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} (1 - S_i) Z_i Y_i^{obs} - \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} (1 - S_i) Z_i Y_i^{fit}(0,0)_{|S_i=0, Z_i=1} \tag{3.14}$$

As we can see, while SvPS approach exploits the SATE, that is information from treated

and control individuals in RCT, BART approach only uses informations from the control group in RCT.
Here the validity of Assumption 3.4 has a crucial role: it ensures that the potential outcome for controls is independent by the indicator of the membership in RCT, for those individuals treated. Thus, the control group in RCT is exchangeable with the treated group in TP: $Y_i(0,0)_{|S_i=0,Z_i=1} = Y_i(1,0)_{|S_i=0,Z_i=1}$, therefore the predicted potential outcome $Y^{fit}(0,0)_{|S_i=0,Z_i=1}$ places the role of the imputed potential outcome $Y_i(0,0)_{|S_i=0,Z_i=1}$.

3.4 Performances of SvPS and BART estimators

In this Section we show the performance of SvPS estimators and BART estimators introduced in previous sections to evaluate the external validity of RCT results and to estimate the PATT from RCT results.
We propose several scenarios, some for which RCT results have external validity and some for which they don't. We conduct the analyses simulating datasets of size: $N = 50\,000$ and $N = 10\,000$. In all scenarios simulations consists of nine covariates (X_{k_i}, $k = 1 \dots 9$) with distributions described in Table 3.1; the sampling indicator, S_i, follows a logistic distribution described in Table 3.2, such that RCT sample size is around 2% of the total dataset size; the binary treatment Z_i which we assume randomly assigned with probability equal to 0.5 for individuals in the RCT sample, while we assume individuals in TP all treated; a continuous outcome, Y_i, having normal distribution $N(\mu, \sigma^2)$, with mean μ described in Table 3.2 Column 2 and variance, σ^2, equal to one.
In **Scenario A** we simulate datasets such that we have an heterogeneous distribution of the outcome under treatment between RCT sample and TP. The mean of the outcome distribution for treated individuals in RCT differs by that one of the treated individuals in TP for a constant quantity equal to 1.2. The covariates involved in the sample selection are all predictors for the outcome.
Also in **Scenario B**, we simulate datasets for which the distribution of the outcome under treatment is heterogeneous between the two groups (TP and RCT), but it is not constant because of the interaction term given by the sample indicator, S_i, the treatment indicator, Z_i, and the covariates X_{1_i} introduced in the mean of the outcome, μ_B Table 3.2 Column 2. In this Scenario we simulate the hypothetical situation for which covariate X_{1_i} is associated with both the outcome and the treatment, that is variable X_{1_i} is an effect modifier, but only for those individuals belonging to the RCT.
In the other scenarios, **C-E**, the simulated datasets are such that we have negligible effects of the sample mechanism on the outcome distribution for treated, that is effect $\tau_{(S,1)}$ is negligible. Focusing on scenarios C and D, the difference between these two scenarios relies on the relationship between covariates involved in the sample mechanism model and those in the outcome distribution. While in **Scenario C** *part* of the covariates involved in the sample mechanism are predictors for the outcome, in **Scenario D** *all of* the covariates involved in the sample mechanism are predictors. In addition, in Scenario

D we introduce an interaction between Z_i and covariate X_{1_i} in the mean of the outcome μ_D, Table 3.2 Column 2, that is like in Scenario B, X_{1_i} is an effect modifier, but in this case also for individuals in the TP. In all scenarios **A-D**, we assume the knowledge of the sample mechanism, that is a correct specification of the PS model.
In order to show how the SvPS approach is sensitive to the correct specification of the PS, we propose other two scenarios, scenario **D-bis and E**, where we suppose a priori a misspecification of the PS. Specifically, in scenario D-bis we estimate the PS, excluding, from the related model, covariates X_{1_i} and X_{2_i} which are also involved in the mean of the outcome distribution. In scenario E we estimate the PS excluding from the model two covariates, X_{1_i} and X_{3_i}, which do not have any relation with the outcome distribution.
For each scenario we simulate 1 000 datasets according with the distributions of the covariates $(X_{k_i}, k = 1 \dots 9)$, the outcome (Y_i), the sample indicator (S_i) and the treatment indicator (Z_i) described above. We fix at 5 the number of strata, and we choose the level of the confidence equal to 95%. We apply the procedure introduced in Section 3.3.1, for checking the negligible effect $\tau_{(S,1)}$ after a primary checking of the well balancing of the covariates between the RCT group and the TP group, within the strata. Figure 3.1 displays, over the one thousands draws, the standardized difference of the means/prevalences of the covariates between the two groups, Austin (2009), for the all Scenarios proposed.
As we can see, in scenarios for which the PS is misspecified, we have that the standardized difference of the means/prevalences for those covariates excluded by the PS model is higher than the others.

Table 3.3 shows the results obtained by SvPS approach to estimate $\tau_{(S,1)}$ and its $C.I._{95\%}(\tau_{(S,1)})$, for the different dataset sizes. In details, for each scenario in Table 3.3 are reported: the true value of $\tau_{(\mathrm{S},1)}$ (Column 3); the bias of the estimator $\hat{\tau}^{\mathrm{SvPS}}_{(S,1)}$(Column 4); the estimated variance of $\hat{\tau}^{\mathrm{SvPS}}_{(S,1)}$, $\hat{\sigma}^2_{\hat{\tau}^{\mathrm{SvPS}}_{(S,1)}}$, (Column 5); the Mean Square Error (MSE) (Column 6); one of the interval estimate calculated among the 1 000 simulated datasets (Column 7).
Looking at the bias (Table 3.3, Column 4) we can observe that it is not high among the scenarios where the PS is correctly specified, less than 0.2, and strictly close to zero in Scenario A. Also the MSE among these scenarios is never high, it is around 0.02. For those scenarios where the PS is misspecified we have different results. As we expected, in scenario D-bis, where the PS is incorrectly estimated an the covariates excluded by the model are predictors for the outcome, we have that the bias is higher than that one calculated in scenario D, with a correct estimate of the PS.
In scenario E, even if the PS is incorrectly estimated, since that the excluded covariates are not predictors of the outcome, the bias is low.
These results suggests that SvPS estimator provides estimate of $\tau_{(\mathrm{S},1)}$ with low bias, under a correct estimate of the PS.
For Scenarios A-D we also apply BART approach to check the generalizability of the RCT results, we use function *bart*, of *BayesTree* package of the R software and we keep the default parameters, in order to predict outcome $Y_i^{fit}(1,1)$ for those individuals

treated in the TP. For simplicity, we report in Table 3.4, with analogue structure to Table 3.3, results obtained by BART approach for dataset size equal to 10 000. Looking at the bias of the estimator $\hat{\tau}^{\text{BART}}_{(\text{S},1)}$ we can observe it is lower than that one of the estimator introduced in SvPS approach, $\hat{\tau}^{\text{SvPS}}_{(\text{S},1)}$, in each scenario. Also BART approach provides estimate of $\tau_{(\text{S},1)}$ with low bias.
In Scenario C, D, D-bis and E the interval estimate for $\tau_{(\text{S},1)}$ covers zero, therefore we can conclude that with confidence level of 95% we have a negligible effect $\tau_{(\text{S},1)}$: the results obtained in RCT can be extended to the TP.
For these last scenarios, we apply, therefore, the stratification approach described in Section 3.3.2 to estimate the PATT for each one of the 1 000 simulated datasets.
Results are summarized in Table 3.5: in Column 2 is reported the true PATT; in Column 4 is reported the bias of the estimator $\hat{\tau}^{\text{SvPS}}_{\text{PATT}}$; in Column 5 is reported the estimated variance, $\hat{\sigma}^2_{\hat{\tau}^{\text{SvPS}}_{\text{PATT}}}$; and finally, in Column 6 it is reported the Mean Square Error (MSE) of the estimator proposed.
Results are also reported in Figure 3.2 which shows the box plot of the estimated PATT values for the 1 000 draws of Scenario C and D, it also displays the true PATT with a black point at 5 for Scenario C, at 7.5 for Scenario D and D-bis and at 5 for Scenario E. The bias of the estimator $\hat{\tau}^{\text{SvPS}}_{\text{PATT}}$, Table 3.5 Column 4, is higher in Scenario D-bis than in the others, it is around 0.30, this shows how this estimator is sensitive to the correct specification of the PS if the covariates involved in the model sampling are also predictors of the outcome. The MSE, Table 3.5 Column 6, in scenarios C-D and E is strictly close to the variance of the estimator, $\sigma^2_{\hat{\tau}_{\text{PATT}}}$. Results obtained suggest that SvPS approach also provides estimate of τ_{PATT} with negligible bias under a correct specification of the PS model.
Finally, we apply BART approach to estimate the PATT in scenario C and D. For semplicity also these last analyses are conducted for dataset size equal to 10 000. Results are displayed in Table 3.6, which has a similar structure to Table 3.5. As we can see the absolute value of the bias of the estimator $\hat{\tau}^{\text{BART}}_{\text{PATT}}$, Table 3.6 Column 4, is strictly close to that one of the estimator $\hat{\tau}^{\text{SvPS}}_{\text{PATT}}$ in both scenarios C-D.

3.5 Medicare data, analysis of simulated data

In this section we conduct analysis on simulated Medicare data, according to a motivating problem of cardiology. Patients with high risk of aortic stenosis are often not candidate for surgical replacement because of the increased risk of operative complications or death. In such patients as alternative it is suggested to apply a less invasive treatment. Here, we want to assess the effect of the Core-Valve Trans-Catheter Aortic Valve Replacement (TAVR) versus Surgical Aortic Valve Replacement (SAVR), on days until first complication (our primary end point) in the TP defined by treated individuals with high risk of aortic stenosis, registered in Medicare database.
Medicare data collects informations of individuals in the US who are aged ≥ 65 years, individuals with disabilities aged < 65 years, and individuals with end-stage renal disease.

Medicare database includes anagraphic informations like gender, age, race and health informations.
In order to reach our goal we use results from a simulated clinical trial, simulated miming one of the trial of Corevalve Network. The trial counts 1184 participants randomized to receive the TAVR vs SAVR. Here, we suppose to have no information about the criteria of enrollment of individuals. The TP counts 21369 individuals of Medicare dataset who received the TAVR.
We first apply the SvPS and BART approaches proposed in Section 3.3.1 to check if there exists any heterogeneity on the outcome distribution under treatment between the TP and the randomized clinical trial. Given the absence of information about the enrollment criteria in the trial, here we estimate the PS excluding the covariates with low prevalence (less than 5%).
Figure 3.3 shows, for each covariate, the standardized difference between the means, prevalences, in the two groups RCT and TP; Pre-stratification process (blue points) and Post-stratification process (orange points) while in Table 3.7 we report the means of the covariates with a low prevalence respect to the indicator of the trial membership, RCT with $S_i = 1$ and TP with $S_i = 0$. As we can see, post-stratification the standardized differences decrease.
The estimated causal effect $\hat{\tau}^{\text{SvPS}}_{(\text{S},1)}$ is equal to 0.758, that is less than one day difference, the $C.I._{.95\%}(\tau_{(\text{S},1)})$ is [-0.199, 1.714] and it covers zero. The $C.I._{.95\%}(\tau_{(\text{S},1)})$ defines a time lapse of around 0-2 days not so wide, this suggests a negligible effect of the indicator of the trial membership, S_i, on the primary outcome. The estimated variance, $\hat{\sigma}^2_{\hat{\tau}^{\text{SvPS}}_{(\text{S},1)}}$ is 0.119, not so high.
Similar conclusions are provided by BART approach. Using estimator $\hat{\tau}^{\text{BART}}_{(\text{S},1)}$ the estimated causal effect is 0.107 with $C.I._{.95\%}(\tau_{(\text{S},1)}) = [-0.155, 0.314]$. Results provided by both approaches suggests negligible effect of the indicator of the trial membership, S_i, that is no significant heterogeneity on average for the outcome distribution under treatment between the two groups.
We than apply the SvPS approach and BART approach, described in Section 3.3.2, to estimate the PATT using RCT data. The estimated PATTs using the two approaches do not differ so much, they are respectively, 3.091 and 2.504. The outcome of interest measures the number of days until first occurrence of a complication, results suggest a positive effect of the TAVR vs SAVR on it. On average those who receive the TAVR shows a first occurrence of complication 2/3 days later than those who receive SAVR.

3.6 Conclusions

In this work we aimed to assess the PATT exploiting the RCT results. We first focus on evaluating the external validity of the RCT results, introducing the causal effect $\tau_{(S,1)}$. The causal effect $\tau_{(S,1)}$ allows us to learn about any heterogeneity of the outcome distribution between the RCT and the TP under treatment. The heterogeneity may be due to several factors as: a RCT not representative of the TP or a different version of

the treatment between the two groups (RCT and TP) irrespective of creating a RCT group whose well matches the characteristics of the TP.
A significant heterogeneity on the outcomes distribution under treatment, between TP and RCT, suggests the absence of external validity of RCT results.
We adopted two approaches, SvPS and BART, providing two estimators for $\tau_{(S,1)}$ and two for τ_{PATT} under the validity of G.Ass. The two approaches differ in terms of procedure and informations used, but produce similar results.
SvPS approach needs a correct PS model specification, in case for which the sample mechanism is not known to the researcher, and it exploits the all informations provided by the RCT. BART approach needs a correct model specification for the outcome of interest and it exploits partially the informations provided by the RCT.
The SvPS estimators, $\hat{\tau}^{\text{SvPS}}_{(S,1)}$ and $\hat{\tau}^{\text{SvPS}}_{\text{PATT}}$, produce estimates of the quantities of interest with low bias as shown in the simulation studies discussed in Section 3.4. Also BART estimators, $\hat{\tau}^{\text{BART}}_{(S,1)}$ and $\hat{\tau}^{\text{BART}}_{\text{PATT}}$ produce estimates of the quantities of interest with low bias, weakly lower than those ones of SvPS approach's estimators.
All of the G.ass. are important in order to identify the PATT from the RCT results. Nevertheless, it would be interesting to consider like a valuable topic for future research the departure from Assumption 3.1 under which the RCT sample is well representative of the TP given a set of pre-treatment variables.

Table 3.1: Distribution of the six covariates: five continuous with normal distribution and one dichotomous following Bernoulli distribution.

$X_{1_i} \sim N(5,2)$	$X_{4_i} \sim N(0,1)$	$X_{7_i} \sim Be(p=0.70)$
$X_{2_i} \sim N(60,2)$	$X_{5_i} \sim Be(p=0.65)$	$X_{8_i} \sim Be(p=0.90)$
$X_{3_i} \sim N(2,1)$	$X_{6_i} \sim N(0,1)$	$X_{9_i} \sim Be(p=0.10)$

Table 3.2: The probability of the sample selection and the mean of the outcome distribution for scenarios A-D.

SCENARIO	$Pr(S_i = 1 \mid \mathbf{X}_i) = p_s$ and Mean of the outcome distribution $Y_i \sim N(\mu, 1)$
A	$p_{(s,A)}(\mathbf{X}_i) = \frac{exp(0.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.4X_{3_i} - 0.3X_{4_i} + 0.3X_{5_i})}{1 + exp(0.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.4X_{3_i} - 0.3X_{4_i} + 0.3X_{5_i})}$ $\mu_A = -3.85 + 1.2Z_i * S_i + 0.5X_{1_i} - 2X_{2_i} - 0.5X_{3_i} + 2X_{4_i} + 2X_{6_i} + 2Z_i$
B	$p_{(s,B)}(\mathbf{X}_i) = \frac{exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i})}{1 + exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i})}$ $\mu_B = -3.85 + 5Z_i + 0.5Z_i * S_i * X_{1_i} + 0.1X_{2_i} - 0.5X_{3_i} + 2X_{4_i} + 2X_{6_i}$
C	$p_{(s,C)}(\mathbf{X}_i) = \frac{exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i} + 0.3X_{5_i})}{1 + exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i} + 0.3X_{5_i})}$ $\mu_C = -3.85 + 5Z + 0.5X_1 - 2X_2 - 0.5X_3 + 2X_4 + 2X_6 + 0.10S * Z$
D	$p_{(s,D)}(\mathbf{X}_i) = \frac{exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i})}{1 + exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i})}$ $\mu_D = -3.85 + 5Z_i + 0.5Z_i * X_{1_i} - 2X_{2_i} - 0.5X_{3_i} + 2X_{4_i} + 2X_{6_i} + 0.10S_i * Z_i$
E	$p_{(s,E)}(\mathbf{X}_i) = \frac{exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i})}{1 + exp(1.5 + 0.1X_{1_i} - 0.1X_{2_i} + 0.1X_{3_i} + 0.1X_{4_i})}$ $\mu_E = -3.85 + 5Z_i - 2X_{2_i} + 2X_{4_i} + 2X_{6_i} + 0.10S_i * Z_i$

Table 3.3: Performance of the SvPS approach to check the generalizability of the results from RCT to the TP. Simulation studies A-E.

SCENARIO	External Validity	$\tau_{(S,1)}$	Bias	$\hat{\sigma}^2_{\hat{\tau}^{\text{SvPS}}_{(S,1)}}$	MSE	$C.I._{95\%}(\tau_{(S,1)})$
For dataset size 50 0000						
A	No	1.2	0.012	0.020	0.020	(0.606, 1.990)
B	No	2.50	0.114	0.007	0.020	(2.141, 3.009)
C	Yes	0.101	0.119	0.006	0.019	(-0.122, 0.757)
D	Yes	0.101	0.116	0.0069	0.0204	(-0.294, 0.728)
D-bis	Yes	0.101	1.012	0.034	1.058	(0.632, 1.797)
E	Yes	0.101	0.099	0.006	0.016	(-0.127, 0.669)
For dataset size 10 0000						
A	No	1.2	0.025	0.121	0.169	(0.676, 3.210)
B	No	2.50	0.121	0.052	0.067	(1.095, 4.094)
C	Yes	0.101	0.120	0.064	0.078	(-1.653, 1.463)
D	Yes	0.101	0.123	0.052	0.067	(-1.302, 1.696)
D-bis	Yes	0.101	1.017	0.192	1.226	(-0.089, 2.893)
E	Yes	0.101	0.099	0.0389	0.049	(-0.911, 1.417)

Table 3.4: Performance of BART approach to check the generalizability of the results from RCT to the TP. Simulation studies A-D.

SCENARIO	External Validity	$\tau_{(S,1)}$	Bias	$\hat{\sigma}^2_{\hat{\tau}^{\text{BART}}_{(S,1)}}$	MSE	$C.I._{95\%}(\tau_{(S,1)})$
For dataset size 10 0000						
A	No	1.20	0.004	0.025	0.025	(1.116,1.521)
B	No	2.50	0.065	0.019	0.023	(2.332, 2.720)
C	Yes	0.101	0.061	0.024	0.027	(-0.148,0.242)
D	Yes	0.101	0.069	0.018	0.023	(-0.130,0.256)

Table 3.5: Performance of the SvPS to estimate the PATT from RCT.

SCENARIO	τ_{PATT}	$\hat{\tau}^{\text{SvPS}}_{\text{PATT}}$	Bias	$\sigma^2_{\hat{\tau}^{\text{SvPS}}_{\text{PATT}}}$	MSE
For dataset size 50 0000					
C	5	4.412	0.104	0.051	0.062
D	7.5	7.952	0.124	0.065	0.077
D-bis	7.502	8.048	0.303	0.090	0.182
E	5.002	5.031	0.105	0.034	0.045
For dataset size 10 0000					
C	5	4.962	0.088	0.459	0.467
D	7.5	7.540	0.106	0.363	0.374
D-bis	7.502	7.565	0.286	0.460	0.541
E	5.002	4.860	0.078	0.261	0.266

Table 3.6: Performance of BART approaches to estimate the PATT from RCT.

SCENARIO	τ_{PATT}	$\hat{\tau}^{\text{BART}}_{\text{PATT}}$	Bias	$\hat{\sigma}^2_{\hat{\tau}^{\text{BART}}_{\text{PATT}}}$	MSE
C	5.002	4.812	-0.071	0.022	0.027
D	7.5	7.431	-0.059	0.019	0.023

Table 3.7: Prevalence of covariates excluded by the PS model, Pre-stratification and Post-stratification for those participants in RCT and those in the TP.

Covariates	Pre-Stratification		Post-Stratification	
(with prevalence <0.05)	S=0	S=1	S=0	S=1
Cancer	0.033	0.036	0.033	0.033
Dementia	0.017	0.014	0.017	0.011
Depression	0.030	0.028	0.030	0.030
Fibrosis-Chronic	0.010	0.012	0.010	0.012
Funct-Dis	0.012	0.018	0.012	0.022
Hyper-Dis	0.014	0.008	0.014	0.008
Hx-MI	0.051	0.052	0.051	0.051
LiverDis	0.010	0.014	0.010	0.012
Maj-Cancer	0.027	0.026	0.027	0.027
Metas-CA	0.005	0.003	0.005	0.005
PCMalnut	0.014	0.013	0.014	0.012
ParkinsonHuntington	0.009	0.008	0.009	0.008
PsychDis	0.010	0.013	0.010	0.011
Seizure disorder	0.009	0.009	0.009	0.009
Severe Hematological	0.006	0.008	0.006	0.008
Sub-Abuse	0.015	0.017	0.015	0.018
Trauma	0.033	0.033	0.033	0.034
Un-Angina	0.042	0.044	0.042	0.046
Vertebral Fractures	0.009	0.007	0.009	0.006

Figure 3.1: Box plots of standardized difference of the means, in 1 000 draws, for each covariate, between the RCT group and the TP group, for the all scenarios.

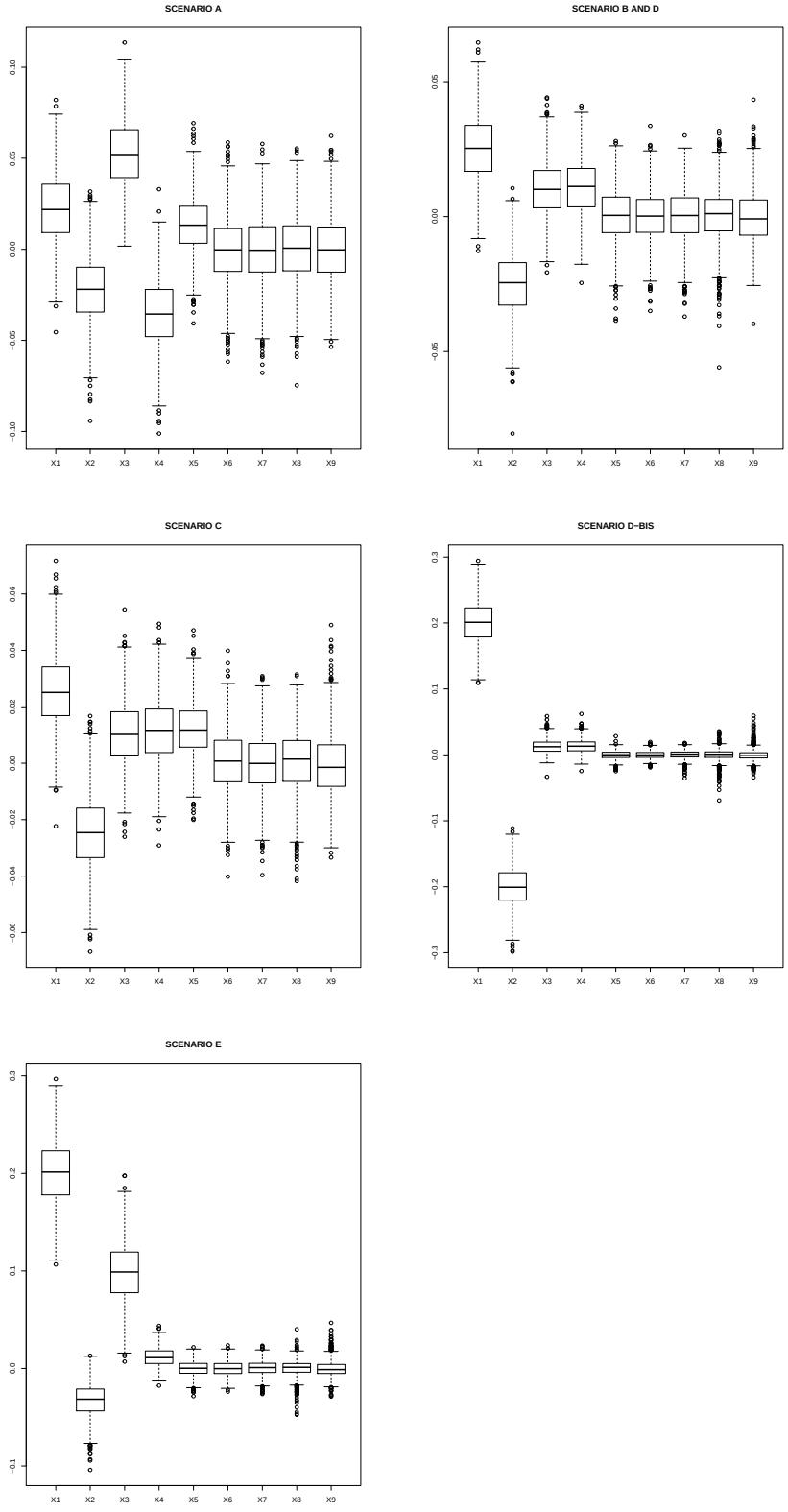


Figure 3.2: Boxplot of the estimated PATT values, in 1 000 draws, using SvPS approach, for: Scenario C (on the top left), Scenario D (on the top right), Scenario D-bis (on the bottom left) and Scenario E (on the bottom right). The true PATT in all scenarios is shown by the black points.

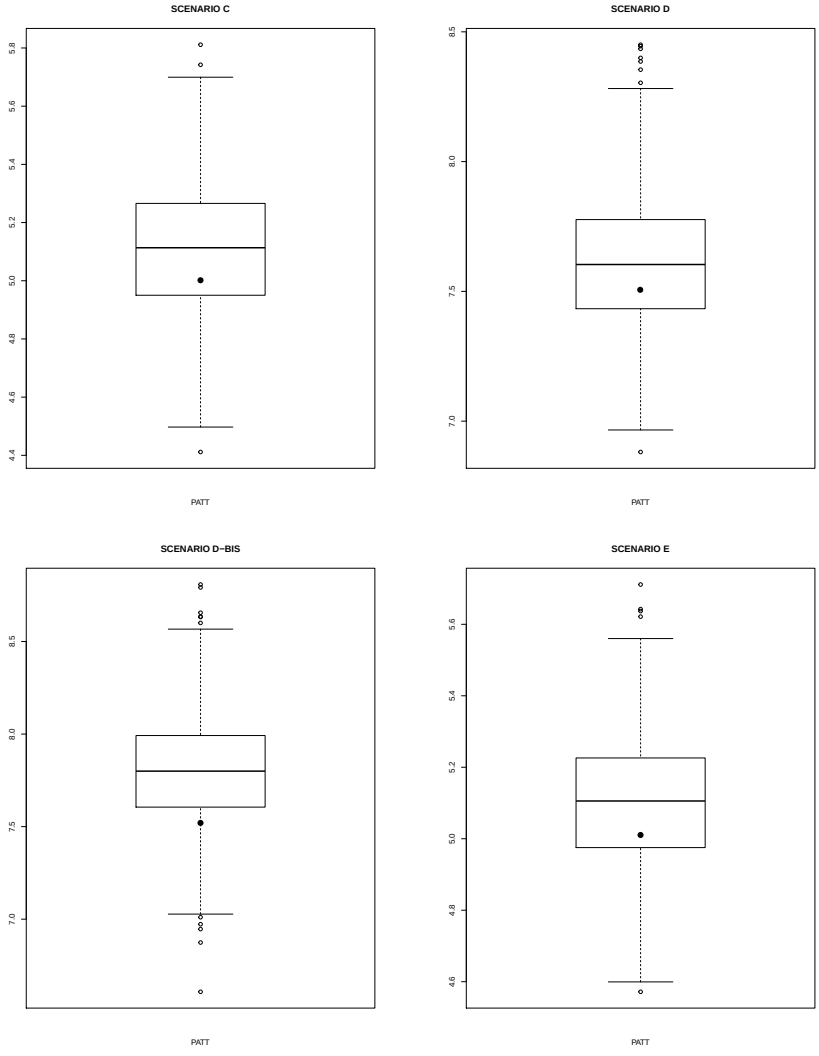
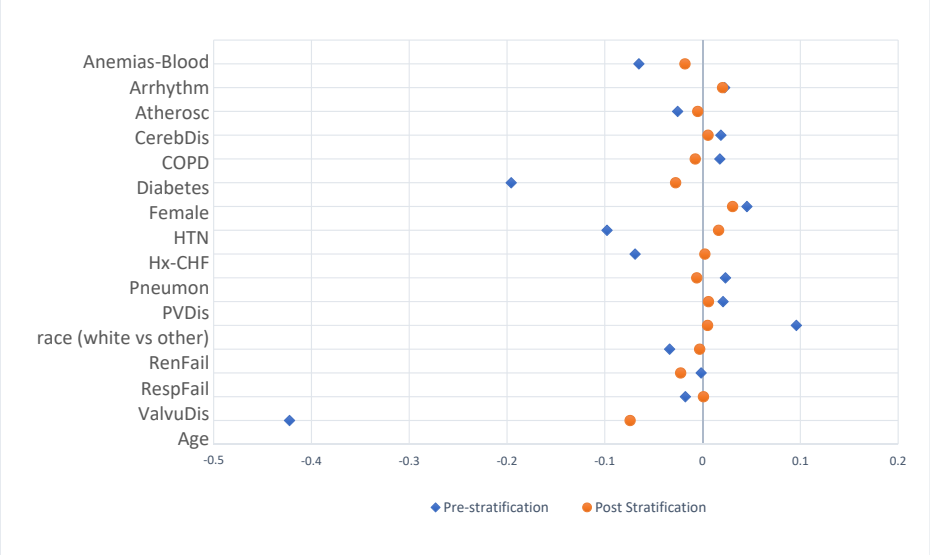


Figure 3.3: Standardized difference of the means/prevalences between the two groups, RCT and TP, for covariates of Medicare Data included in the PS model.



Acknowledgements

I want to say thanks at first to my tutor Alessandra, because she supported me during these last years (which are not only three) *directly and indirectly*. Thanks also for the weekends spent working with/helping me. I tried to do my best even if sometimes I have to admit that I wanted to give up.
Many thanks also to the others professors with whom I worked during these last three years: Michela, Cory and Joe.
Thanks to my Family and my friends, you are a lot and I do not want to exclude anyone making a list. Please forgive me if I mention in particular Vale, Sere and Giulia (the mare!)
Thanks to Kico for the times he reminded me who I am, thanks for being simply as you are.
Thanks to my beautiful colleagues and ex-colleagues even if *I am not one of you* expect for my Guru. You supported me at work a lot, but in particular Jules (the investigator) and her favourite colleague (who is not me) who helped me a lot before my visiting in Boston.
Thanks to Giuseppe and professor Claudio who allowed me to work at CISIA in smart-working mode during these years.
Three full years of research and life experiences. I realised most of the dreams I had since I was a child.
All this thanks to you.

Federica

Bibliography

E.A. Akl, A.D. Oxman, J. Herrin, G.E. Vist, I. Terrenato, and F. Sperati. Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst Rev*, 2011.

S. Athey and G.W. Imbens. The state of applied econometrics - causality and policy evaluation. *ArXiv working paper*, No 1607.00699, 2016.

P.C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med.*, 28(25): 3083–3107, 2009. doi: 10.1002/sim.3697.

M. Baccini, L. Ghirardi, D. Farinella, and A. Biggeri. Comparison of two statistical indicators in communicating epidemiological results to the population: a randomized study in a high environmental risk area of italy. *BMC Public Health*, 19:733, 2019. doi: https://doi.org/10.1186/s12889-019-7003-y.

E. Battistin and E. Rettore. Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, 142: 715–730, 2008.

S. M. Berry and D. A. Berry. Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60:418–426, 2004.

M.D. Cattaneo and J. C. Escanciano. Regression discontinuity designs: Theory and applications. *Advances in Econometrics*, 38. Emerald Group Publishing. To appear, 2016.

M.D. Cattaneo, B. R. Frandsen, and R. Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 3(1):1–24, 2015.

H.A. Chipman, E.I. George, and R.E. McCulloch. Bart: Bayesian additive regression trees. *The annals of Applied Statistics*, 4(1):266–298, 2010.

S.R. Cole and E.A. Stuart. Generalizing evidence from randomized clinical trial to target populations the actg 320 trial. *American Journal of Epidemiology*, 172(1):107–115, 2010. doi: 10.1093/aje/kwq084.

A. Diamond and J.S. Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95(3):932–945, 2013.

D. Farinella, G. Terni, A. Biggeri, and M. Baccini. Communicating epidemiological results through alternative indicators: cognitive interviewing to assess a questionnaire on risk perception in a high environmental risk area. *Cogent Social Sciences*, 3(1), 2017. doi: https://doi.org/10.1080/23311886.2017.1319539.

R.A. Fisher. *The Design of Experiments*. Oliver and Boyd, London, 1935.

C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

E.D. Hahn and R. Soyer. Probit and logit models: Differences in the multivariate realm. *Available at: http://home.gwu.edu/ soyer/mv1h.pdf*.

J. Hansen, L. Holm, L. Frewer, P. Robinson, and P. Sandøe. Beyond the knowledge deficit:recent research into lay and expert attitudes to food risks. *Appetite*, 41, 2003.

E. Hartman, R. Grieve, R. Ramsahai, and J.S. Sekhon. From sate to patt:combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society*, 2015.

K. Hirano, G.W. Imbens, D.B. Rubin, and Zhou X.H. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1:69–88–327, 2000.

K. Imai, D. Tingley, and T. Yamamoto. Experimental design for identifying causal mechanisms. *Journal of Royal Statistical Society, Series A*, 176(1):5–51, 2013.

G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635, 2008.

G. W. Imbens and D. B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25:305–327, 1997.

W.G. Imbens and D.B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction*. Cambridge University Press, New York, NY, USA, 2015.

D. S. Lee. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697, 2008.

D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 485:281–355, 2010.

C.R. Lesko, A.L Buchanan, D. Westreich, J.K. Edwards, M.G. Hudgens, and S.R. Cole. Generalizing study results. *Epidemiology*, 28:553–561, 2017.

F. Li, A. Mattei, and F. Mealli. Bayesian inference for regression discontinuity designs with application to the evaluation of italian university grants. *The Annals of Applied Statistics*, 9(4):1906–1931, 2015.

J.K. Lunceford and M. Davidian. Stratification and weighting via propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23:2937–2969, 2004. doi: 10.1002/sim.1903.

A. Mattei and F. Mealli. Application of the principal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics*, 63:437–446, 2007.

A. Mattei and F. Mealli. Regression discontinuity designs as local randomized experiments. *Observational Studies*, 2:156–173, 2016.

R. Nethery, F. Mealli, and F. Dominici. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *Annals of Applied Statistics*, 2019.

S.M. Perkins, W. Tu, and et al. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 9:93–101, 2000.

P.R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:597–610, 2019.

P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P.R. Rosenbaum and D.B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79: 516–524, 1984.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. B. Rubin. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6:34–58, 1978.

D. B Rubin. Discussion of "randomization analysis of experimental data in the fisher randomization test" by basu. *Journal of the American Statistical Association*, 75: 591–593, 1980.

D.B. Rubin. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statistical Science*, 21(3): 299–309, 2006.

G. Salanti, AE. Ades, and JP. Ioannidis. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol.*, 64:163-71., 2011.

A. Sales and B. B. Hansen. Limitless regression discontinuity: Causal inference for a population surrounding a threshold. *ArXiv working paper*, No 1403.5478, 2015.

L. Sjöberg. Factors in risk perception. *Risk Analysis*, 20:1–12, 2000.

C.R. Smith, M.B. Leon, M.J. Mack, and D.G. Miller. Transcatheter versus surgical aortic-valve replacement in high-risk patients. *The New England Journal of Medicine*, 364, 2011. doi: 10.1056/NEJMoa1103510.

E.A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25:1–21, 2010. doi: 10.1214/09-STS313.

E.A. Stuart, S.R. Cole, C.P Bradshaw, and P.J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A*, 174: 369–386, 2001. doi: 10.1111/j.1467-985X.2010.00673.x.

M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

D. Thistlethwaite and D. Campbell. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960.

J. Z. Zhang and D. B. Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.

D.A. Zipkin, C.A. Umscheid, N.L. Keating, E. Allen, K. Aung, and Beyth R. et al. Evidence-based risk communication: a systematic review. *Ann Intern Med.*, 161, 2014.