UNIVERSITY OF TRIESTE

Department of Mathematics, Informatics and Geosciences
PhD Course in Applied Data Science and Artificial Intelligence
Cycle XXXVIII

Statistical modeling approaches for uncertainty reduction

Candidate: Lea Anna Cozzucoli

Supervisors: Prof. Francesco Pauli

Co-supervisors: Prof. Angelos Alexopoulos, Prof. Andrea Tracogna

20 February 2026

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVIII CICLO DEL DOTTORATO DI RICERCA IN

APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Finanziato dall'Unione europea - NextGenerationEU
Funded by the European Union - NextGenerationEU

STATISTICAL MODELING APPROACHES FOR UNCERTAINTY REDUCTION

Settore scientifico-disciplinare: **STAT-01/A Statistica**

DOTTORANDO / A
LEA ANNA COZZUCOLI

COORDINATORE
PROF. FRANCESCO PAULI

SUPERVISORE DI TESI
PROF. FRANCESCO PAULI

CO-SUPERVISORE DI TESI
PROF. ANGELOS ALEXOPOULOS
PROF. ANDREA TRACOGNA

ANGELIS ALEXOPOULOS
28/12/2025 23:58

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | UNIVERSITÀ DEGLI STUDI DI TRIESTE

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Francesco Pauli, for his invaluable help, continuous support, and kindness throughout these years. Along with him, my deepest gratitude goes to my co-supervisors: Prof. Angelos Alexopoulos, for his constant presence, guidance, and dedication during my visiting period and beyond, and to Prof. Andrea Tracogna for his support and encouragement.

A special thank you goes to the MIB Trieste School of Management, which not only funded my scholarship but, more importantly, made me feel truly part of its community and provided me with the opportunity to meet wonderful people with whom I shared this journey and grew both personally and professionally.

A truly heartfelt thank you goes to Prof. Gioia Di Credico for her immense patience, kindness, guidance, constant availability and support during this journey. Her many amazing (and colorful) ideas shine brightly throughout these pages.

I would also like to warmly thank Prof. Valeria Edefonti, University of Milan, for her helpful suggestions and insightful guidance.

A huge thank you goes to all the statisticians at DEAMS, University of Trieste, with whom I have shared this journey from the very beginning. You quickly became a second family, making everything feel lighter. Each of you has enriched me in a unique way, and while it would deserve far more beautiful words, this space is simply not enough to fully express my gratitude. A very special thank you goes to Vincenzo, for the amazing lunches that set very high standards for our lunch breaks, and for his help with practically everything.

To all the professors I encountered along this path, as well as to the colleagues from the PhD with whom I shared this incredible journey, thank you. But my warmest and most heartfelt thanks goes to my friend Gino, with whom I shared the joys and the sorrows of this path from the first exams, through the sleepless nights at SUS always having fun, despite the many breakdowns. Thanks to you, this journey was much lighter and full of unforgettable moments.

I cannot move forward without thanking Prof. Ioannis Ntzoufras, the entire Department of Statistics at Athens University of Economics and Business, and especially all the

members of the AUEB Computational and Bayesian Statistics Lab, who welcomed me in a way I had never experienced before and made my visiting period truly unforgettable. In particular, I would like to thank Argyro and Anna for making me feel part of the group from the beginning and for all the amazing moments we shared together. Thanks to you, Athens truly feels like home to me.

Finally, I want to thank my family and friends and all of those whose constant presence carried me through this journey. Above all, I thank my dad, to whom I owe everything, whose loving presence gave me the strength to move forward and the courage to never let go of life and, my boyfriend Roberto, who walked beside me through both light and difficult days, sharing every step with immense patience and love. You are not only the people I love most, but also the scientists and statisticians I look up to and hope to resemble a little more each day. Thank you for all the love you give me; you are the heart of everything I do.

Last, but certainly not least, I thank my grandparents. Even though you are no longer physically with me, your immense love continues to guide me through life, and I will never be able to thank you enough for all that you have given me.

Contents

List of Figures

List of Tables

Introduction

Statistical modeling nowadays represents the core of empirical research across various scientific disciplines. Ranging from various scientific disciplines, statistical models are used to extract meaningful insights from data (Breiman, 2001). Nevertheless, statistics has brought notable changes to scientific research, enabling to mathematically represent real-world phenomena in order to study them; however, such representations may not be fully correct. In fact, as Box, 1979 famously noted, "*All models are wrong, but some are useful*," since they are merely approximations of real-world systems and therefore only approximately correct. This simple phrase encapsulates both the greatest strength and the main limitation of statistical techniques. While diving into the world of statistics, uncertainty emerges as a pervasive and unavoidable element, extending well beyond the boundaries of statistical modeling itself. As De Finetti, 2017 noted:

> In almost all circumstances, and at all times, we all find ourselves in a state of uncertainty. Uncertainty in every sense. Uncertainty about actual situations, past and present.

In this sense, uncertainty is not merely a technical challenge, but an intrinsic feature of human existence, reinforcing the close connection between statistical reasoning and the way people navigate reality. Seen through this lens, uncertainty is neither something that can be avoided nor a limitation to be eliminated; rather, it should be understood and quantified in order to be properly addressed within statistical reasoning.

The philosophical foundations of statistical inference have always concentrated around the nature and sources of uncertainty. In the frequentist tradition, build upon ideas of Fisher, Pearson and Neyman, a major focus was put on sampling variability, recognizing that finite samples yielded to estimates that oscillated around true population values (Fisher, 1925; Neyman and Pearson, 1933). Following that, the Bayesian inferential

approaches offered in fact a framework able to quantify estimating uncertainty and at the same time incorporate prior knowledge assuming that the parameter we are actually estimating is an aleatory variable and that therefore yields uncertainty with itselfs. Generally, all the conclusions drawn from any statistical analysis are necessarily subject to multiple sources of uncertainty (Chatfield, 1995). Understanding and possibly reducing this uncertainty represents one of the central challenges in modern statistics. Generally, uncertainty arises from data as well from and so the sources may be classified as evidenced as: (i) uncertainty about the structure of the model, (ii) uncertainty about estimates of the model parameters and, (iii) unexplained random variation in observed data.

In this thesis three different projects are developed that address different source of uncertainty across three distinct methodological and applied contexts. While the specific methods and applications differ across chapters, they share the core idea behind that is to provide more robust, transparent, and well-calibrated statistical conclusions. The thesis is organized in three chapters each presenting an original contribution to statistical field with applications to real-world problems.

In the first chapter of this thesis a Bayesian dynamic factor model for causal inference in panel data settings with staggered treatment adoption is proposed. The model captures temporal dependence through latent factors evolving according to first order auto-regressive processes and extends to non-Gaussian outcomes via an exponential family formulation. A key methodological innovation is the proposed individual treatment effect estimator, defined as the difference between posterior predictive means under treated and control sub-models, which integrates out idiosyncratic noise to yield lower-variance estimates with well-calibrated credible intervals. The methodology is validated through simulations across Gaussian, Negative Binomial, and Bernoulli outcomes, demonstrating improvements in accuracy and interval quality. Subsequently, an empirical application to macroeconomic data explores the effect of targeted audits on firm compliance behavior, revealing heterogeneous treatment responses consistent with the reduction of fraud behaviors.

The second chapter, explain the development of a graphical framework inspired by multiverse analysis to systematically evaluate how preprocessing and modeling decisions affect epidemiological risk estimates. Rather than committing to a single analytical path, the approach estimates models across all defensible specification combinations and visualizes the distribution of results to assess robustness. Specifically, we studied the relationship between tobacco smoking, alcohol consumption, and upper aerodigestive tract cancer risk across 1,944 model specifications. Results demonstrate that while

overall conclusions remain qualitatively robust, quantitative estimates can meaningfully depend on specification choices, particularly the treatment of extreme values in data preprocessing.

Finally, then in the third and last chapter a Bayesian Additive Regression Trees model is employed to estimate heterogeneous treatment effects of artificial intelligence and machine learning usage on sustainability practices adoption among European small and medium-sized enterprises. Using European survey data the analysis examines effects on recycling, resource consumption reduction, energy efficiency, and sustainable product development relying on an adequate estimator that can capture variations of causal effects across some population subgroups. Results provide robust evidence of positive causal effects across all sustainability practices, with substantial heterogeneity by firm size and geography.

Chapter 1

Bayesian causal inference for observational panel data

1.1 Introduction

Assessing the possible presence of causal relationships is becoming crucial in many scientific fields ranging from, economics (Abadie *et al.*, 2010, Hsiao *et al.*, 2012) and epidemiology to social sciences (Ben-Michael *et al.*, 2023). Our point of departure in this work is that researchers often cannot rely exclusively on randomized trial data, even though such trials are the conceptual ideal for assessing the causal impact of an intervention, since treatment assignment is unbiased by construction. Consequently, relying on observational data, such as administrative records (Abadie *et al.*, 2010) or electronic health files (Stuart *et al.*, 2013), is a common practice in many studies. This has led to the development of causal inference techniques that can recover the treatment effect unbiasedly, despite the potentially confounded nature of the data (Imbens and Rubin, 2015). In panel data structures, the outcome of interest is observed for every unit at multiple time points (both before and after the treatment) and, thus, causal analysis has to deal with additional challenges: (i) presence of time-varying unobserved confounding, (ii) unit level heterogeneity, (iii) temporal dependence of the outcome, (iv) staggered treatment adoption, (v) limited treated cases. Traditional two-way fixed effect (TWFE) or difference-in-differences (DiD) estimators are usually employed and they handle only a subset of the exposed issues. In particular, they rely on the parallel trend assumption that, in the absence of treatment, the average outcome of treated and control units would have evolved equally (Abadie, 2005). Violations of this assumption, driven either by time-varying confounders or unit-specific trends, can severely bias causal

estimates. To overcome this problem a possible solution is to condition upon covariates. Anyway, this approach does not guarantee complete control over all confounders thus, a potential bias cannot be entirely excluded.

A seminal method that does not require the parallel trends assumption for the identification of causal effects is the so-called Synthetic Controls (SC) technique developed by Abadie and Gardeazabal, 2003 and Abadie *et al.*, 2010. In SC, the counterfactual for the treated is inferred by optimally weighting untreated units sharing similar characteristics. The method was originally developed for a single treated unit and is often applied sub-optimally when multiple treated units are present. Moreover, the method, is particularly suited in the cases when just a few treated units are available and may control for both observable and hidden time-varying confounders making causal inference assumption more plausible. In general, the SCM is a particular case of a latent factor model (see Appendix in Abadie *et al.*, 2010), in which untreated outcomes depend on a small number of time-varying factors and unit-specific loadings. Thus, by taking a weighted combination of control units, a synthetic unit is constructed that should replicate the latent structure of the treated unit taking into account unobserved heterogeneity and thus not requiring the parallel trends assumption. However, SC methods present several limitations. Notably, it can be applied only to a single outcome at a time, and it does not explicitly model the temporal correlation between repeated measurements of the outcome. Furthermore, when the number of units is small, SC struggles to provide valid inference for the estimated causal effects. Building on this framework, Gobillon and Magnac (2016), see also Xu (2017) and Athey *et al.* (2021),developed the generalized synthetic control (GSC) method, which extends the SC approach to accommodate multiple treated units and formally estimates the latent factors and unit-specific loadings. GSC enables to explicitly model unobserved, time-varying heterogeneity through a small set of common factors with unit-specific loadings. These latent factors capture complex baseline trends and shocks, offering a flexible framework for causal inference with panel data. Although factor-based estimators improve robustness to hidden biases, they still exhibit some limitations. Importantly, for most of the recent approaches it is hard to obtain uncertainty intervals for the causal effects of interest especially when the interest is on the estimation of unit-specific causal effects. To address this limitation, we propose a Bayesian causal inference method that naturally allows for uncertainty quantification of treatment effect. Building on the latent factor analysis framework, we develop a dynamic model where latent factors evolve according to an autoregressive process of the first order (AR(1)), naturally capturing temporal dependence and accommodating staggered adoption. Specifically, the main contribution of this approach lies in three

key aspects: (i) A Bayesian estimation procedure is proposed that treats all unknown quantities probabilistically, enabling to easily provide uncertainty quantification for both average and unit-specific treatment effects, even when the number of treated units is small. Unlike frequentist GSC, the proposed Bayesian approach yields exact posterior credible intervals directly from the posterior distribution. (ii) The model is extended to non-Gaussian outcomes by proposing an exponential-family formulation that unifies multiple likelihoods under a common latent factor structure. This generalization relaxes the normality assumption that is common to the most factor based causal estimators, allowing the model to handle different types of data through appropriate link functions while preserving interpretability. (iii) An alternative modeling approach for the causal effect is introduced, that represents the main key innovation, by fitting two parallel sub-models (one for the control and one for treated observations). Each sub-model captures its own latent structure and temporal dynamics, and causal effects are then derived as the difference between their posterior predictive means. This formulation explicitly separates the mechanisms underlying treated and untreated units, integrates out idiosyncratic noise, and yields interpretable, unit and time specific treatment effects. Estimation proceeds within the Bayesian paradigm, using an efficient Markov chain Monte Carlo (MCMC) sampler that ensures accurate posterior inference even when the treated group is small. Through extensive simulation studies we demonstrate that the proposed approach can recover the true causal effects across various data generating scenarios and, at the same time, provides well calibrated uncertainty credible intervals, improved precision and enhanced robustness for the quantity of interest estimation.

The empirical advantages of the model are illustrated by analyzing administrative-tax records from the Independent Authority for Public Revenue in Greece, quantifying how audits affect the compliance behavior of firms in a fraud prone sector. The outcome of interest is the daily number of issued receipts with strong overdispersion making it an ideal setting to test our non-Gaussian extensions. The estimated ITEs reveals the presence of a slight heterogeneity in both the magnitude and timing of the response, heterogeneity that would probably be masked by conventional methods. For most firms, the estimated effects are negative, showing a significant drop in the issuance of receipts immediately after the intervention. At the aggregate level, posterior distributions of the average treatment effect remain consistently below zero in the post-audit period, confirming a systematic compliance-improving impact.

1.2 Potential outcome framework

1.2.1 Notation

Let $i = 1, \ldots, N$ index the observational units and $t = 1, \ldots, T$ the time points. For each unit i we observe the outcome y_{it} at every period t, defining the vector $y_{i1}, \ldots, y_{iT}$. In what follows we refer to units that have experienced the treatment (in our case being audited) as the *treated units* while, those that never do as *controls*. To distinguish upon those two categories we define d_{it} as the treatment indicator, $d_{it} = 1$ if unit i has been treated by time t and $d_{it} = 0$ otherwise, and denote by T_i the first period in which unit i is treated. Finally, N_{tr} and $N_c = N - N_{tr}$ are the numbers of treated and control units, respectively.

1.2.2 Causal effect framework

Within the potential outcome framework, for each unit i at each time point t there are two *potential outcomes*, $y_{it}^{(0)}$ and $y_{it}^{(1)}$, that can be observed. As previously stated, only one of these outcomes is actually observable, while the other, known as the *counterfactual*, remains unobserved and must be inferred from the data; for the treated units $y_{it} = y_{it}^{(1)}$ while, $y_{it}^{(0)}$ is the one missing that needs to be estimated. In the case of panel data, the problem is naturally extended to multiple counterfactual; thus, whenever i belongs to the treated set N_{tr} and t exceeds intervention time T_i, the counterfactual $y_{it}^{(0)}$ is missing, and conversely for control units.

In this setting, we propose an estimand for the Individual Treatment Effect (ITE) that builds on the idea of modelling both the treated and untreated outcomes. Rather than defining the effect as the difference between the observed outcome and an estimated counterfactual, we consider the expected difference between the two potential outcomes. Specifically, the individual treatment causal effect $\tilde{\tau}_{it}$ estimator for every unit i and time t is defined as

$$\tilde{\tau}_{it} = \mathbb{E}\left[y_{it}(1)\right] - \mathbb{E}\left[y_{it}(0)\right], \tag{1.1}$$

and can be interpreted as the difference between the expected value observed under treatment and the expected in the opposite case. The estimand of interest integrates out idiosyncratic and measurement error uncertainty, yielding to lower uncertainty and more stable estimations of the causal effect. Conceptually, it represents the average outcome difference for units sharing similar latent characteristics at a given time point. Our estimator differs from classical ITEs estimand used in matrix completion approaches

(Athey *et al.*, 2021), which are typically defined by conditioning on the imputation of the missing counterfactual for treated units,

$$\widehat{\tau}_{it} = y_{it} - \widehat{y}_{it}(0), \qquad i \in N_t,\ t > T_i. \tag{1.2}$$

and represents the realised effect for each treated unit at a specific time $t_i > T_i$. In contrast, $\tilde{\tau}_{it}$ targets the structural, unit and time specific causal effect that comparable units would experience on average, while naturally incorporating uncertainty quantification within the Bayesian framework. Recovering the entire path of the ITEs, denoted by τ_{it}, reveals heterogeneous and dynamic treatment responses that would be obscured by aggregate estimands such as the ATT commonly used in DiD or TWFE. This allows us to study how effects evolve over time, which is particularly relevant in our empirical setting. The unbiased identification of the causal effect τ_{it} is based on the following assumptions:

- **Consistency**: ensures that the treatment is well defined and applied consistently across units, so that no alternative versions of the intervention could lead to different outcomes.
- **Ignorability or Unconfoundedness**: imposes that conditional on observed covariates, the treatment assignment is independent of potential outcomes.
- **Overlap**: guarantees that both treated and untreated units have a non-zero probability of being allocated to either the treatment or control groups and, jointly with the assumption of unconfoundedness, define the strong ignorability
- **Stable Unit Treatment Value Assumption** (SUTVA) imposes no interference between units, implying that potential outcomes for the i-th unit are unrelated to the treatment status of other individuals (Angrist *et al.*, 1996).

In addition to these assumptions, as discussed in the previous section, causal inference in panel data also requires the parallel trends assumption. In the proposed model, this requirement is relaxed since, conditional on the latent structure, all time-varying confounding is accounted for. More details are provided in Appendix A.1.

1.3 Bayesian Dynamic Factor Framework

In this section, we introduce the modeling and inferential framework at the basis of our causal analysis. First, a general latent factor model specification that accommodates

staggered treatment adoption and time dependence is presented. Next, we outline the prior specification and we present the concepts that lies besides our proposed estimator.

1.3.1 Latent Factor Specification and Treatment Partitioning

Suppose that at each time point $t \in \{1, \dots, T\}$ a N dimensional vector $\mathbf{y}_t = \big(y_{1t}, \dots, y_{it}\big)$ is observed. Let $d_{it} \in \{0, 1\}$ denote a binary indicator equal to 1 from the unit specific intervention time $T_i \in \{1, \dots, T, \infty\}$ onward and 0 otherwise. Conditional on natural parameters $\boldsymbol{\theta}_t = (\theta_{1t}, \dots, \theta_{it})$, we assume that the N elements of vector y_t are conditionally independent and that each potential outcome $y_{it}^{(d_{it})}$ follows a probability distribution that belongs to the exponential family

$$f\big(y_{it}^{(d_{it})} \mid \theta_{it}, \phi\big) = \exp\left\{ \frac{T\big(y_{it}^{(d_{it})}\big)\, \theta_{it} - A(\theta_{it})}{a(\phi)} + c\big(y_{it}^{(d_{it})}, \phi\big) \right\}, \tag{1.3}$$

where $T(\cdot)$ is a sufficient statistic, $A(\cdot)$ and $c(\cdot)$ are known functions, and ϕ is a dispersion parameter. The conditional mean of the outcome is therefore

$$\mu_{it}^{(d_{it})} = \mathbb{E}[T(y_{it}^{(d_{it})}) \mid \theta_{it}^{(d_{it})}] = A'(\theta_{it}^{(d_{it})}). \tag{1.4}$$

The treatment indicator $d_{it} \in \mathbf{D}$ induces a non-overlapping partition of the outcome matrix $\mathbf{Y}$ into (i) pure control observations for which $d_{it} = 0$ for all t; (ii) pre-treatment rows of treated units where $d_{it} = 0$ for all $t < T_i$; and (iii) post-treatment observations characterized by $d_{it} = 1$ for $t \geq T_i$. Only block (iii) is affected by the intervention, while blocks (i) and (ii) can inform latent structure of untreated potential outcomes and are used to reconstruct the path of $y_{it}^{(0)}$. The same logic applies while recovering the path of treated potential outcomes using the partition (iii). This motivates modeling two parallel sub-models,

$$\mathcal{M}_C : \{\text{control observations} \quad (i,t) \quad \text{such that} \quad d_{it} = 0\}, \tag{1.5}$$

$$\mathcal{M}_{Tr} : \{\text{treated observations} \quad (i,t) \quad \text{such that} \quad d_{it} = 1\}. \tag{1.6}$$

so that each observation contributes only to the likelihood of the sub-model corresponding with its treatment status ad time t. This structure naturally accommodates staggered adoption. The likelihood then factorizes into two components. Let Θ_C and Θ_{Tr} denote the parameters of $\mathcal{M}_C$ and $\mathcal{M}_{Tr}$, respectively then the full likelihood is therefore the

product of

$$\mathcal{L}_C(\Theta) = \prod_{i=1}^{N}\prod_{t=1}^{T} f(y_{it} \mid \Theta)^{\,1-d_{it}}, \tag{1.7}$$

and

$$\mathcal{L}_{Tr}(\Theta) = \prod_{i=1}^{N}\prod_{t=1}^{T} f(y_{it} \mid \Theta)^{\,d_{it}}. \tag{1.8}$$

The treated and control potential outcomes are modeled using exponential family distributions whose natural parameters is driven by a latent factor structure and observed covariates where available. Specifically, the potential outcomes are modeled as

$$y_{it}^{(0)} \mid \theta_{it}^{(0)} \sim f\big(y_{it}^{(0)} \mid \theta_{it}^{(0)}, \phi\big) \tag{1.9}$$

$$y_{it}^{(1)} \mid \theta_{it}^{(1)} \sim f\big(y_{it}^{(1)} \mid \theta_{it}^{(1)}, \phi\big), \tag{1.10}$$

where the natural parameter is linked to the latent structure via the canonical link function $h(\cdot)$:

$$\theta_{it}^{(d_{it})} = h^{-1}\big(\zeta_{it}^{(d_{it})}\big), \tag{1.11}$$

the linear predictors are defined as:

$$\zeta_{it}^{(0)} = \mathrm{w}_i \mathrm{f}_t^{\top}, \tag{1.12}$$

$$\zeta_{it}^{(1)} = \tau_{it} + \mathrm{w}_i \mathrm{f}_t^{\top}, \tag{1.13}$$

where f_t is the t-th row of the latent factor matrix $\mathbf{F} \in \mathbb{R}^{T\times K}$, w_i is the i-th row of the factor loading matrix $\mathbf{W} \in \mathbb{R}^{N\times K}$, and β is a vector of treatment effect coefficients on observed covariates x_{it}. Given the definition of the Individual Conditional Treatment Effect (ICTE) in (1.1), let $\widehat{\theta}_{it}^{(1)}$ and $\widehat{\theta}_{it}^{(0)}$ denote the posterior natural parameters for treated and control cases respectively. The proposed estimator is then

$$\tilde{\tau}_{it} = A'\big(\widehat{\theta}_{it}^{(1)}\big) - A'\big(\widehat{\theta}_{it}^{(0)}\big), \tag{1.14}$$

which, using (1.4), is equivalently written as

$$\tilde{\tau}_{it} = \widehat{\mu}_{it}^{(1)} - \widehat{\mu}_{it}^{(0)}, \qquad i \in N_{tr},\ t \geq T_i, \tag{1.15}$$

with posterior predictive means

$$\widehat{\mu}_{it}^{(1)} = \mathbb{E}[y_{it} \mid \mathcal{M}_{Tr}, \mathcal{B}], \tag{1.16}$$

$$\widehat{\mu}_{it}^{(0)} = \mathbb{E}[y_{it} \mid \mathcal{M}_{C}, \mathcal{B}], \tag{1.17}$$

where $\mathcal{B}$ denotes the observed data. The ICTE therefore captures the structural difference in posterior predictive means implied by the treated and control latent structures, integrating out idiosyncratic noise and respecting the exponential–family outcome distribution. This formulation highlights that the causal effect arises directly from differences in the latent structures of the treated and control models given the relation in (1.11). Thus, by focusing on the difference between treated and control latent factors, we obtain unit and time specific causal effects that naturally respect the underlying outcome distribution and provide interpretable predictions of the intervention's impact.

1.3.2 Prior specification

Under a Bayesian framework, is essential to specify prior distribution for the parameters of interests. Here, priors for the model parameters $\boldsymbol{\Theta}$ are defined for both control model $\mathcal{M}_{\mathcal{C}}$ and treated model $\mathcal{M}_{\mathcal{T}}\nabla$. Same specification is adopted for both sub-models. We rely on the use of an adequate prior able to capture the temporal dependence in the latent factors to guarantee the dynamic part of the model. Let $\mathbf{f}_t = (\mathrm{f}_{t1}, \ldots, \mathrm{f}_{tk})^\top$ denote the K latent factors at time t and let, $\mathbf{F} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_t^\top) \in \mathbb{R}^{(\mathrm{K}\times\mathrm{T})}$ be the factor matrix. We specifically impose a zero mean Gaussian process prior with exponential decay

$$\mathbf{F} \sim \mathrm{N}(0, \mathbf{C}) \tag{1.18}$$

where $\mathbf{C}$ is a block-diagonal covariance matrix with $\mathbf{C}^{(k)} \in \mathbb{R}^{(\mathrm{T}\times\mathrm{T})}$ blocks so that so that, for $i, j \in \{1, \ldots, T\}$,

$$\left[\mathbf{C}^{(k)}\right]_{ij} = \frac{\sigma_k^2}{1-\rho_k^2}\,\rho_k^{|i-j|}, \qquad \rho_k \in (-1,1), \quad \sigma_k^2 > 0. \tag{1.19}$$

For each block, the covariance matrix $\mathbf{C}^{(k)}$ is precisely the autocovariance structure of the an autoregressive (AR) process of order one $f_{tk} = \rho_k\, f_{t-1,k} + \xi_{tk}$ with $\xi_t \sim \mathcal{N}(0, \sigma^2)$. The block diagonal structure implies that each latent factor evolves as a stochastic model by itself, imposing independence between the K factors. Setting $\rho_k \equiv \rho$ and $\sigma_k^2 \equiv \sigma^2$ for all K, yields a more parsimonious specification with identical covariance blocks $\mathbf{C}^{(k)} = \mathbf{C}_0$. Regarding the loading matrix $\mathbf{W}$, each loading vector $\mathbf{w}_i = (w_{i1}, \ldots, w_{ik})$,

$i = 1, \ldots, N$, is assigned with a standard normal prior $\mathbf{w}_i \sim \mathrm{N}\big(\mathbf{0}, \mathbf{I}_K\big)$. For canonical one-parameter exponential-family outcomes (e.g. Bernoulli, Poisson) no additional prior is needed. But, when a separate dispersion parameter exists, i.e. Gaussian variance, we treat it as an unknown parameter placing a weakly informative Gamma prior (details in Appendix A.2).

1.3.2.1 Identifiability and constraints

The identifiability is a significant issue and challenge in latent factor models, since it is a necessary prerequisite for model estimation and interpretability. For unidimensional latent factors, the identification issue is confined to the fact that the scale of both the loadings and the factors is not uniquely determined. In the multidimensional setting, however, identification is much more complex since any invertible linear transformation of the latent variables, combined with a corresponding adjustment to the loading matrix, leaves the distribution of the observed responses unchanged. A way to overcome this issue is to constrain the loading matrix to be lower triangular (Geweke and Zhou, 1996) or to assuming orthogonal factors and loadings (Bai and Ng, 2006; Fan *et al.*, 2017). In particular, for any orthogonal matrix $Q \in \mathbb{R}^{K \times K}$, the product $WF^\top$ is unchanged under the rotation $W \leftarrow WQ$ and $F \leftarrow FQ$, implying that the individual components of W and F are not uniquely identified, even though the implied low-rank signal is. In this work, identifiability is addressed at the level of the quantities of interest, namely posterior predictive means and treatment effects, which depend on $WF^\top$ and are invariant to such rotations. Finally, the effective complexity of the latent structure is controlled through the choice of K. Since the causal estimand is computed from $\mu_{it}^{(1)}$ and $\mu_{it}^{(0)}$ rather than from individual factor components, remaining non-identifiability of (W, F) does not propagate to the proposed ITE estimator.

1.4 Computational aspects

Inference is conducted via a tailored MCMC sampler that generates draws from the joint posterior distribution $\pi(\mathrm{W}, \mathrm{F}, \phi, \varphi_F | \mathbf{Y}, \mathbf{D})$ where $\mathbf{F}$ and $\mathbf{W}$ denote the latent factors and loadings, ϕ refers to the dispersion parameter (where needed) and φ_F the hyperparameters governing the dynamic prior on the latent factors (AR(1) parameters that are autocorrelation parameter and innovation variance). Given the set of parameter of interest the target posterior distribution is:

$$\pi(\mathrm{W}, \mathrm{F}, \phi, \varphi_F \mid \mathbf{Y}, \mathbf{D}) \;\propto\; \mathcal{L}_C(\Theta)\, \mathcal{L}_{Tr}(\Theta)\, \pi(\Theta), \tag{1.20}$$

where $\mathcal{L}_C$ and $\mathcal{L}_{Tr}$ are the control and treated likelihood components defined in (1.7)–(1.8), and $\pi(\Theta)$ denotes the joint prior over the parameters. Whenever a parameter block admits a closed-form of the full conditional as in the case of the loading matrix $\mathbf{W}$, the variance σ^2 of the AR(1) prior or the outcome dispersion parameter, a Gibbs sampler is used straightforward. For the AR(1) persistence parameter ρ we use random-walk Metropolis–Hastings. Details about this steps are available in Appendix. Particular care is reserved for the latent factor matrix $\mathbf{F}$ where an auxiliary-gradient based Metropolis Hasting is adopted following Titsias and Papaspiliopoulos, 2018. The method combines an elliptical slice proposal with a short, likelihood guided gradient push, thereby exploiting the gradient of the likelihood while remaining exactly invariant to the Gaussian prior. Simulation evidence shows that this sampler delivers higher effective sample sizes than similar algorithm like Metropolis–adjusted Langevin algorithm (MALA) (Roberts and Stramer, 2002) or the pre-conditioned Crank–Nicolson scheme (pCN) (Cotter *et al.*, 2013); moreover, it only requires one tuning parameter that can be set adaptively during burn-in phase. With $\mathbf{F}$ updated in this way, the chain mixes rapidly even in high-dimensional panels. The algorithm steps are described in 2 while, the subsequent subsection explains the underlying augmented target construction that makes the move both non-local and gradient-driven.

1.4.1 Auxiliary gradient sampler for latent factors

Let $g(\mathbf{F}) = \ell(\mathbf{Y} \mid \mathbf{F}, \mathbf{D}, \boldsymbol{\vartheta})$ denote the log-likelihood and let $\mathbf{C} \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}}$ be the prior covariance matrix of $\mathbf{F}$. Following Titsias and Papaspiliopoulos, 2018, we first enlarge the state space with an auxiliary draw:$\mathbf{z} \sim \mathcal{N}\Big(\mathbf{F} + \frac{\delta}{2}\nabla g(\mathbf{F}),\, \frac{\delta}{2}\mathbf{I}\Big)$, which produces the following expanded target

$$p(\mathbf{F}, \mathbf{z} \mid \mathbf{Y}, \mathbf{D}, \boldsymbol{\vartheta}) \;\propto\; e^{\,g(\mathbf{F})}\, \mathcal{N}(\mathbf{F} \mid \mathbf{0}, \mathbf{C})\, \mathcal{N}\Big(\mathbf{z} \mid \mathbf{F} + \tfrac{\delta}{2}\nabla g(\mathbf{F}),\, \tfrac{\delta}{2}\mathbf{I}\Big). \tag{1.21}$$

A first order Taylor expansion of $g(\cdot)$ around the current $\mathbf{F}$ shows that, *conditional on* $\mathbf{z}$

$$q(\mathbf{F}^* \mid \mathbf{z}) = \mathcal{N}\Big(\mathbf{F}^* \;\Big|\; \tfrac{2}{\delta}\mathbf{Q}\,\mathbf{z},\, \mathbf{Q}\Big), \qquad \mathbf{Q} = \big(\mathbf{C}^{-1} + 2\mathbf{I}/\delta\big)^{-1}, \tag{1.22}$$

is an independent proposal that still exploits $\nabla g(\mathbf{F})$. The Metropolis–Hastings statistic simplifies to

$$log(\alpha) = g(\mathbf{F}^*) - g(\mathbf{F}) + \nu(\mathbf{z}, \mathbf{F}^*) - \nu(\mathbf{z}, \mathbf{F}), \tag{1.23}$$

$$\nu(\mathbf{z}, \mathbf{F}) = \big(\mathbf{z} - \mathbf{F} - \tfrac{\delta}{4}\nabla g(\mathbf{F})\big)^{\top} \nabla g(\mathbf{F}), \tag{1.24}$$

since all normalizing constants depending only on $(\mathbf{z}, \boldsymbol{\vartheta})$. Titsias and Papaspiliopoulos (2018) prove that marginalizing (1.21) over $\mathbf{z}$ produces a chain that dominates its auxiliary counterpart in the Peskun sense, so any ergodic average enjoys lower asymptotic variance than the random walk or MALA alternatives. We adapt δ during the burn-in to achieve an acceptance rate of 50–60 % as suggested for this type of algorithms; moreover, the chain for $\mathbf{F}$ typically mixes an order of magnitude faster than with MALA or pCN. Once the eigen-decomposition of $\mathbf{C}$ is done, generating $\mathbf{F}^*$ and evaluating the acceptance ratio costs $\mathcal{O}(n^2)$ per update, which is negligible relative to likelihood evaluation for the panels considered here. This auxiliary–gradient step completes the latent-factor update. The detailed algorithm steps are described in Appendix A.3.

Algorithm 1 Auxiliary–gradient based update for the latent factors $\mathbf{F}$

Require: Data $\mathbf{Y}$; treatment indicator $\mathbf{D}$; current state $(\mathbf{F}, \boldsymbol{\vartheta})$; step sizes δ

1: Draw an auxiliary variable: $\mathbf{z} \sim \mathcal{N}\big(\mathbf{F} + \frac{\delta}{2}\nabla g,\ \frac{\delta}{2}\mathbf{I}\big)$
2: Draw a proposal candidate: $\mathbf{F}^*$ following (1.22)
3: Compute Metropolis–Hastings statistic as in (1.23) and (1.24)
4: Draw $u \sim \text{Uniform}(0, 1)$; **if** $\log u < \log \alpha$ **then** $\mathbf{F} \leftarrow \mathbf{F}^*$
5: **return** updated $\mathbf{F}$

(Adapt δ to reach ≈50–60 % acceptance.)

1.5 Simulation study

To evaluate the performance of the proposed model, we conduct a series of simulation studies across different data types and likelihood functions. In particular, we focus on assessing the model's efficiency in recovering the causal effects in settings involving continuous, count, and binary outcomes with temporal dependencies. Simulating data under controlled conditions allows us to systematically examine the model behavior in terms of bias, variance, and overall estimation accuracy. To highlight the advantages of estimating $\tilde{\tau}_{ti}$ in (1.1) over $\hat{\tau}_{ti}$ in (1.2), we simulate, for each time t and unit i, the potential outcomes $y_{ti}(0)$ and $y_{ti}(1)$ from (1.12) and (1.13) respectively.

Synthetic panel data are generated under different scenarios for all considered distributional cases. In each case, the conditional mean of the outcome is linked to the latent component through its canonical link function: identity for continuous, log for count, and logit for binary outcomes. Treatment adoption is staggered in all cases, and outcomes are driven by a common two-factor latent structure. Unit and time specific treatment effects are drawn independently from a normal distribution $\text{N}(0, \sigma_\tau^2)$, capturing heterogeneous treatment effects across units and time. The three distributional cases considered differ

TABLE 1.1: Selected values for the parameters of interest in the simulated data.

Scenario	σ_0^2	σ_1^2	ψ
S1	1.0	1.0	0.0
S2	1.0	2.0	0.5
S3	0.5	1.5	0.9
S4	1.2	1.2	0.4

only in their outcome specific likelihood functions; the latent factor specification and treatment assignment mechanism remain the same. Performance is assessed on panels with $(T, N) = (36, 200)$ dimensions under varying combinations of unit variances σ_0^2, σ_1^2 and their correlation $\psi_{\sigma_0^2,\sigma_1^2}$. Among the N units, 40 are treated with staggered adoption; for each treated unit N_{tr} the treatment onset T_i is drawn randomly but constrained to occur no earlier than period $t = 10$. After generating the data and estimating the models, we compare the recovered causal effect trajectories of $\tilde{\tau}_{ti}$ and $\hat{\tau}_{ti}$ using the following metrics: (i) root mean square error (RMSE), (ii) the mean width of the 95% credible interval, (iii) the empirical coverage, and (iv) the mean interval score (MIS). For each scenario, we run 20,000 MCMC iterations with a 10,000 burn-in using the sampling strategy described in Section 1.4. The step size parameter δ is adaptively tuned during burn-in to achieve an acceptance rate of 50–60%. Convergence for all parameters is confirmed by trace plots. We keep fixed across all simulation designs the latent factor process, the staggered treatment assignment mechanism, and the heterogeneous treatment effect. Only the outcome distribution and its canonical link function vary between scenarios. This allows us to assess how the proposed estimator $\tilde{\tau}_{ti}$ performs under different likelihoods while controlling for all other features of the data-generating process. The scenario tested across the three likelihood use combinations $(\sigma_0^2, \sigma_1^2, \psi)$ in Table 1.1.

1.5.1 Continuous cases

For continuous responses we assume a Gaussian likelihood with identity link. Specifically, we assume

$$y_{it}^{(d_{it})} \mid \theta_{it}^{(d_{it})} \sim \mathcal{N}\big(\theta_{it}^{(d_{it})},\, \sigma_y^2\big), \tag{1.25}$$

so that the conditional mean coincides with the natural parameter. The treated and control predictors follow (1.12)–(1.13) where the identity link directly connects the mean to the latent structure.

In the simulation study, the latent factors $\mathbf{f}_t = (f_{t1}, \ldots, f_{tK})$ are generated as independent realizations from multivariate normal distributions with AR(1)-type covariance

matrices C_k across time, as defined in (1.19). The factor loadings $\mathbf{w}_i$ are sampled independently from standard normal distributions, and the latent component is constructed as the matrix product $\mathbf{F}\mathbf{W}^\top$. The treatment effect intercept τ_{it} is sampled independently for each unit i and time t from a normal distribution, and enters the treated predictor as a unit- and time-specific shift.

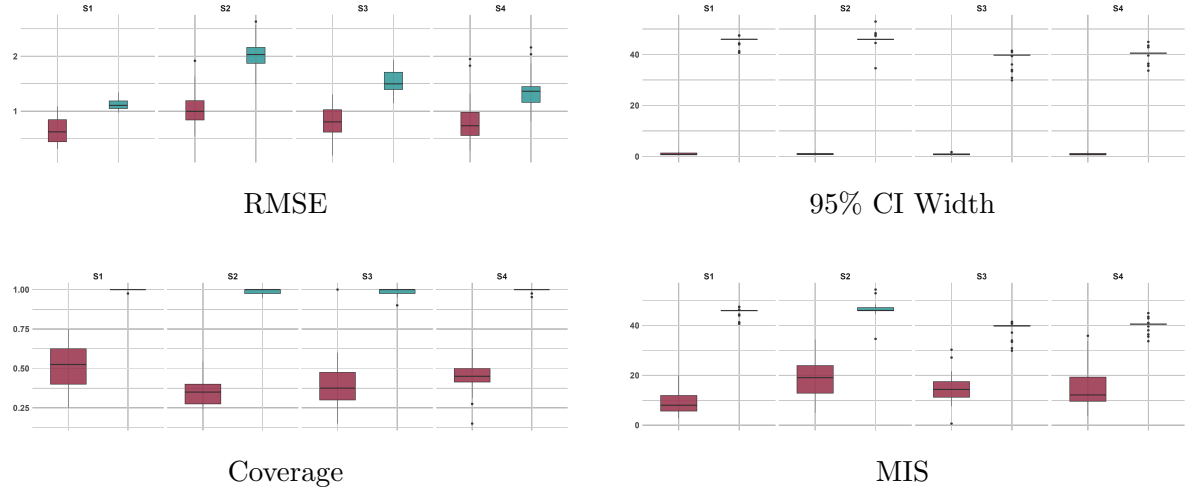


FIGURE 1.1: Comparison of ITE estimates across scenarios S1–S4 for the Gaussian case. Red denotes the proposed mean-based ITE $\tilde{\tau}_{ti}$; blue denotes the estimator $\hat{\tau}_{ti}$. The four panels display (top row, left to right) RMSE and credible-interval width, and (bottom row, left to right) empirical coverage and Mean Interval Scores (MIS). Each panel shows boxplots for all four scenarios.

1.5.2 Count cases

For count responses with overdispersion, we adopt a Negative Binomial (NB) likelihood with a log link,

$$y_{it}^{(d_{it})} \mid \theta_{it}^{(d_{it})} \sim \mathrm{NB}\Big(\mu_{it}^{(d_{it})},\, \phi\Big)\,, \qquad \mu_{it}^{(d_{it})} = \exp\big(\theta_{it}^{(d_{it})}\big), \tag{1.26}$$

where $\mu_{it}^{(d_{it})}$ is the conditional mean and $\phi > 0$ is a fixed dispersion (shape) parameter. Under the canonical log link $h(\mu) = \log \mu$, the natural parameters are again determined by the latent predictors in (1.12)–(1.13). In the data-generating mechanism, we retain the same latent factor structure and treatment design as in the Gaussian case, with the mean entering the Negative Binomial rate via the log link.

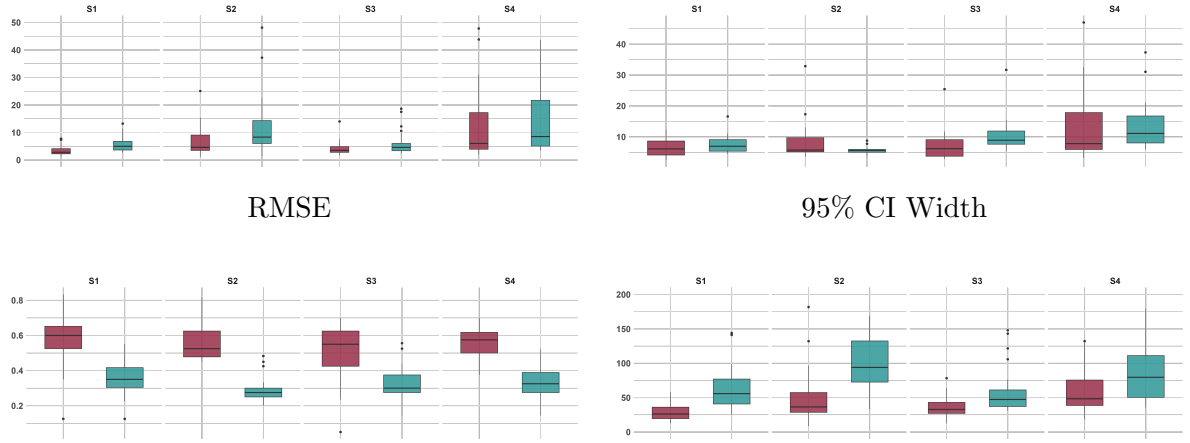


FIGURE 1.2: Comparison of ITE estimates across scenarios S1–S4 for the Negative Binomial case. Red denotes the proposed mean-based ITE $\tilde{\tau}_{ti}$; blue denotes the estimator $\hat{\tau}_{ti}$. The four panels display (top row, left to right) RMSE and credible-interval width, and (bottom row, left to right) empirical coverage and Mean Interval Scores (MIS). Each panel shows boxplots for all four scenarios.

1.5.3 Binary cases

For binary outcomes we use a Bernoulli likelihood with logit link,

$$y_{it}^{(d_{it})} \mid \theta_{it}^{(d_{it})} \sim \text{Bernoulli}\Big(p_{it}^{(d_{it})}\Big), \qquad p_{it}^{(d_{it})} = \text{logit}^{-1}\big(\theta_{it}^{(d_{it})}\big). \tag{1.27}$$

The natural parameters are obtained from the same latent predictors described in (1.12)–(1.13), and the simulation setup mirrors that of the previous cases, while the probability $p_{it}^{(d_{it})}$ is obtained by applying the logistic transformation to the corresponding latent predictor.

1.5.4 Simulation results

In all simulation panels, the red boxplots represents our mean-based ITE estimator (1.1), obtained by fitting treated and control sub-models, while the blue ones are showing results for $\hat{\tau}_{ti}$ from (1.2), computed as the observed outcome minus the estimated counterfactual at each post-treatment time. Across Gaussian, Negative Binomial, and Bernoulli outcomes, the proposed estimator closely recovers the ITE paths, reduces variability, and improves precision by integrating out idiosyncratic noise. These results appear as consistently lower RMSE and better MIS without excessively failing in terms of coverage. The lower MIS is driven by narrower credible intervals, that are most evident in the Gaussian case (Figure 1.1), which exhibits the smallest uncertainty even

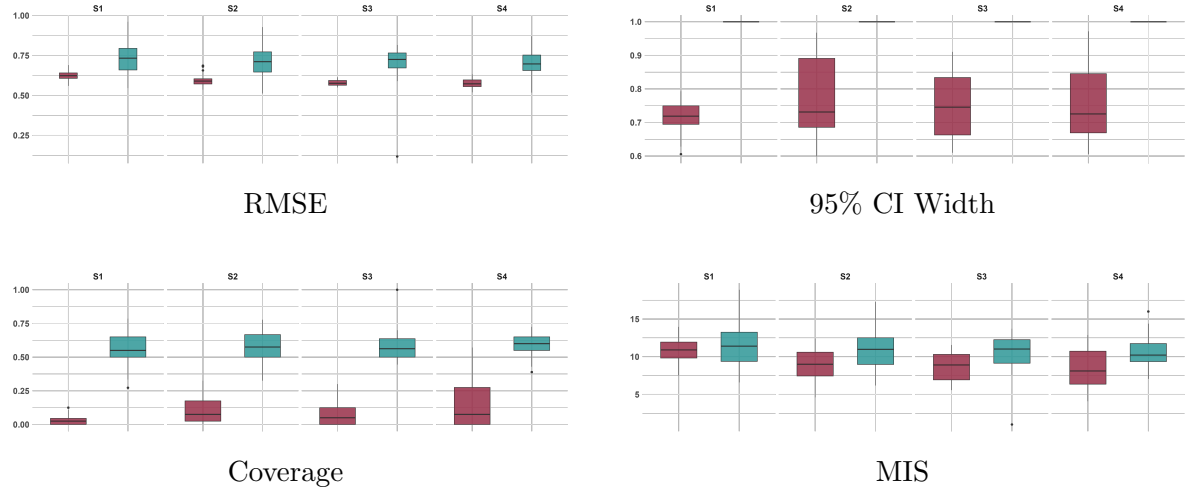


FIGURE 1.3: Comparison of ITE estimates across scenarios S1–S4 for the Bernoullian case. Red denotes the proposed mean-based ITE $\tilde{\tau}_{ti}$; blue denotes the estimator $\hat{\tau}_{ti}$. The four panels display (top row, left to right) RMSE and credible-interval width, and (bottom row, left to right) empirical coverage and Mean Interval Scores (MIS). Each panel shows boxplots for all four scenarios.

though this implicitly leads to a higher coverage for the blue estimator. For counts data (Figure 1.2), intervals are wider due to mean–variance coupling and link-function curvature, yet RMSE and MIS still favor the proposed estimator. Moreover, in this case our estimator achieves also higher coverage levels. Regarding Bernoullian data (Figure 1.3 the proposed estimator is still performing better in terms of MIS even though the difference is less evident than in other case. Comparing scenarios in Table 1.1, we can see that as ψ (treated–untreated correlation) increases, the two sub-models share more common variation and the red and blue curves get a little bit closer; while, when ψ is smaller, differences become more pronounced. In all cases, the proposed estimator $\tilde{\tau}_{ti}$ remains more stable and achieves better accuracy and interval quality with respect to $\hat{\tau}_{ti}$. Detailed time-series plots of metrics across all post-treatment periods are provided in Appendix A.4.

1.6 Empirical application

1.6.1 Data and Context

The application focuses on the causal effect of targeted audits on daily reporting behavior among N businesses in the Greek energy sector, an industry with documented fraud caused by under reporting and receipt manipulation. The audit program prioritized firms suspected of receipt inflation, characterized by the systematic release of large

numbers of low-value (micro) receipts to inflate reported transaction counts and hide off-invoice (untaxed) quantities sold to customers. Our outcome is represented by the count of receipt y_{ti} issued on day t by business i. Within the potential outcome framework we define the ITE as $\tau_{ti} = y_{ti}(1) - y_{ti}(0)$ where treatment d_{ti} is defined as $d_{ti} = 1$ once the business is being audited and $d_{ti} = 0$ otherwise. Conceptually, the main idea is that the audit (intervention) should reduce the issuance of fake micro-receipts causing a post-audit decline in y_{it}. To identify the ITE we leverage the dynamic factor model introduced in Section 1.3 modeling time-varying and unit level heterogeneity via latent factors, relaxing the parallel trends assumption conditional on the latent structure. Our empirical work contributes to strengthen the literature documenting compliance responses to audits, complementing prior works with clear evidence on the heterogeneity of post-audit effects (Kleven *et al.*, 2011; DeBacker *et al.*, 2018; Advani *et al.*, 2023; Christiansen, 2024). Figure 1.4 shows the raw outcome of treated and untreated observations by time,

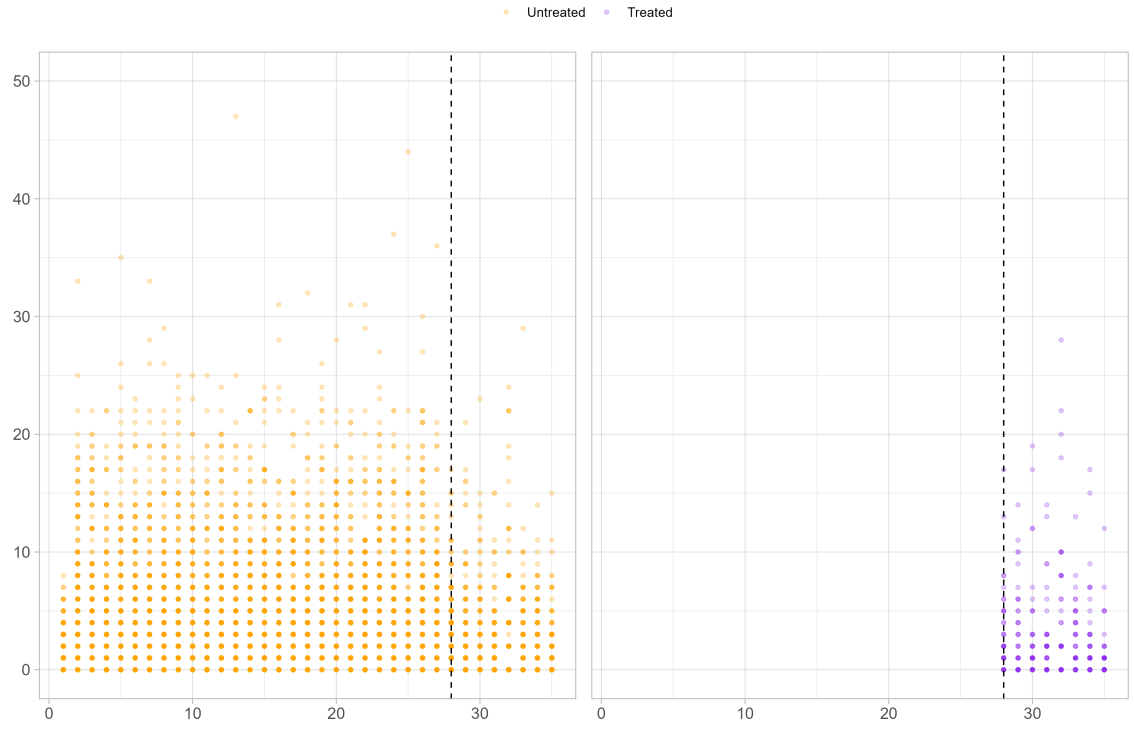


FIGURE 1.4: Daily receipts by calendar date, split by current treatment status. Left panel shows untreated units at date t ($d_{ti} = 0$) while, right panel shows treated at date t ($d_{ti} > 0$). Each dot is a firm daily observation (t, i) with y_{ti} receipts. Dashed line shows the time point at which units are being audited T_i.

separating observations by current treatment status at each time point t. The sample data collection goes from 1 November 2024 till 4 December 2024 and includes $N = 200$ firms; 38 firms become treated at $t = 28$ so, from 27 November 2024. Visually, treated

points cluster at lower receipt counts compared to untreated observations at the same timeline, also, overall dispersion seems to be more compressed after audits, suggesting a reduction in receipt inflation among audited firms. To complement the visual evidence a dispersion index $\text{DI} = \text{Var}(y)/\mathbb{E}(y)$ is evaluated to check if the counts vary more than a basic Poisson model would expect. DI measured on treated, controls and overall data is in all cases well above 1, indicating presence of overdispersion (values are approximately around 5), justifying the Negative Binomial specification for the counts.

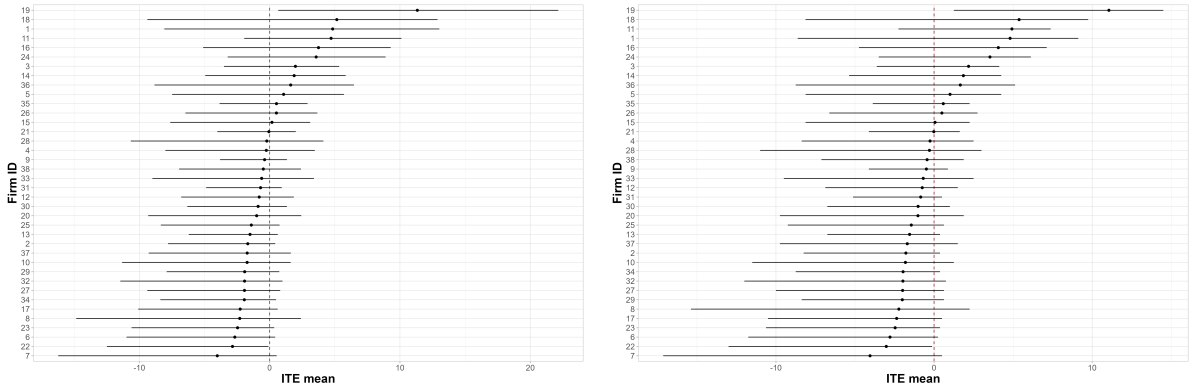
1.6.2 Model fitting and posterior inference

We fit the Bayesian dynamic factor model described in Section 1.3, estimating treated/control sub-models and subtracting their conditional means to obtain $\tilde{\tau}_{it}$ (our proposed mean-based ITE estimator). This approach conditions on unobserved time-varying confounding using a low-rank latent structure with AR(1) dynamics; exact posterior credible intervals follow from MCMC computation as described in Section 1.4. Model convergence diagnostics such as ESS, $\hat{R}$ and traceplots are analyzed to confirm the goodness-of-fit of the model over the data.

1.6.3 Results

In terms of results, since we suppose that audits should primarly suppress the number of micro receipt, rather than the real demans, we expect a reduction of y_{it} after T_i concentrated in a lower amount of receipts and also a reductions in terms of distribution over-dispersion. So, regarding the ITE we expect to have negative ITE's meaning that what the firms declare after audit is a lower number of receipt that they would if they were not audited.

Figure 1.5 summarizes firm-level ITEs aggregated over post-audit days for the proposed estimator (panel a) and for the classical one (panel b). Consistent with the mechanism that audits suppress micro receipt production, rather than genuine demand, most posterior means lie on the negative side, indicating fewer receipts post-audit relative to the counterfactual without audit. At the same time, the distribution remain heterogeneous: many firms exhibit clearly negative effects, while a subset exhibits a positive effect and for some the effect appears to be null. A notable number of ITEs shows a 95% credible interval that intersects zero. This could be due to the short post-audit window (five days) or to the discrete and overdispersed nature of counts, each of which widens intervals at the unit level. Comparing panels, $\tilde{\tau}$ deliver tighter intervals than $\hat{\tau}$, reflecting the fact that the mean-based estimator integrates out idiosyncratic noise through the



(a) (b)

FIGURE 1.5: Posterior estimates with 95% credible intervals of firm-level ITE averaged over post-treatment periods. Points are posterior means; horizontal bars are 95% credible intervals; the vertical dashed line marks zero. Panel (a) shows ITE based on posterior predictive means ($\tilde{\tau}$), while panel (b) shows ITE based on posterior predictive draws ($\hat{\tau}$).

treated/control sub-models and tend to produce thinner intervals. This closely aligns with the simulation evidence indicating that the proposed approach improves precision without sacrificing coverage.

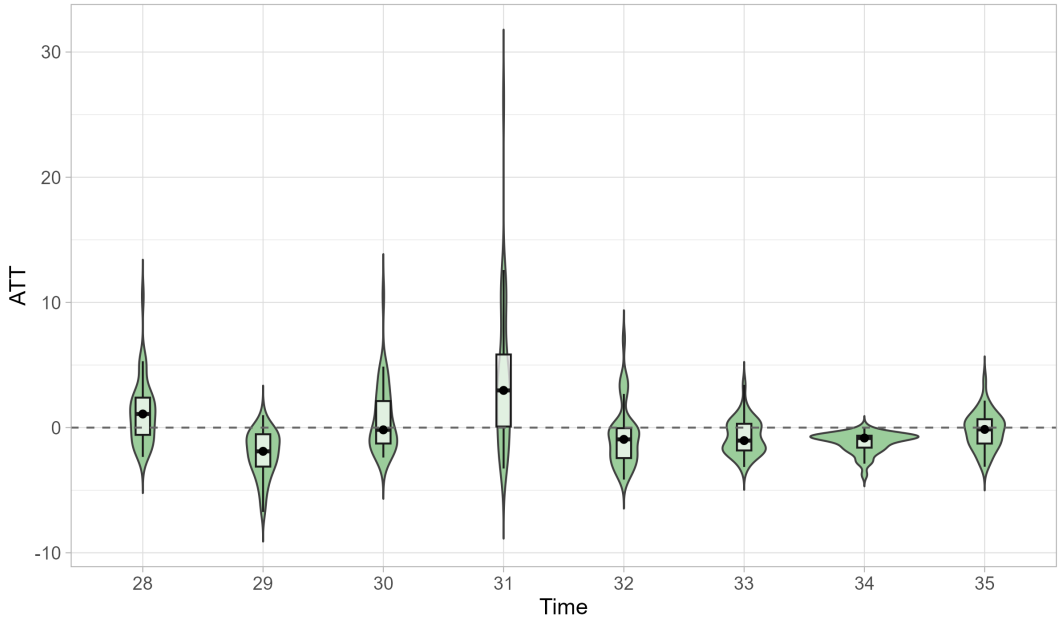


FIGURE 1.6: Posterior distribution of the daily average treatment effect of the treated (ATT) by post-audit day. Each violin shows the across-firm distribution of the mean ITE at day t; the embedded box gives the median and interquartile range, and the dot marks the posterior mean. The dashed horizontal line indicates zero.

Figure 1.6 displays the posterior distribution of the daily average treatment effect over the post-audit period. On most dates the location of the distribution is at or below zero,

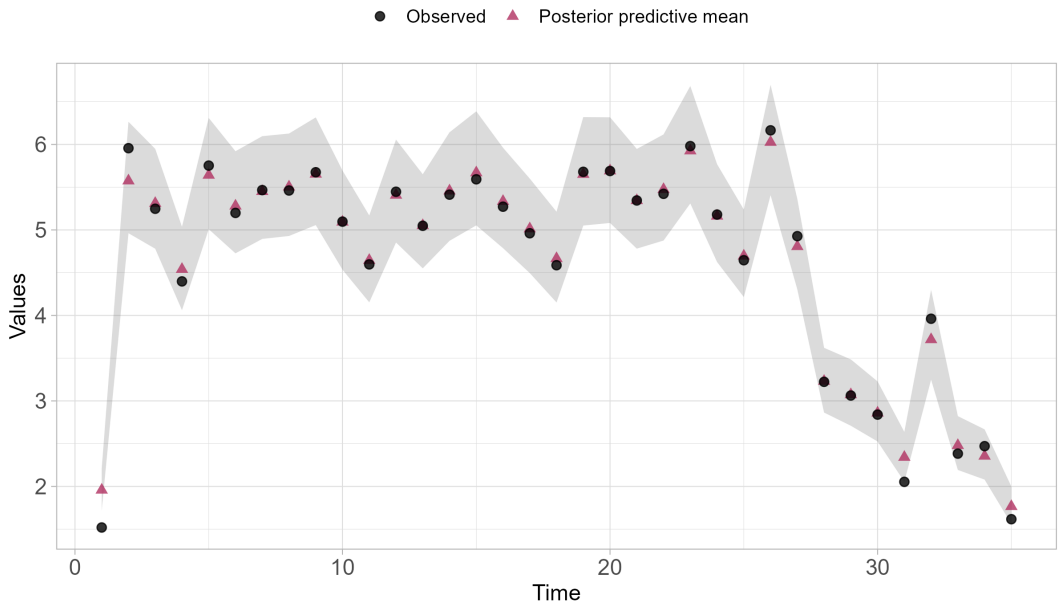


FIGURE 1.7: In-sample fit of daily receipts over the full sample period. The grey ribbon shows 95% posterior predictive intervals; black dots are posterior predictive means; red triangles are observed receipts. For $t < T_i = 28$ predictions come from the control sub-model $\mathcal{M}_c$, and for $t \geq T_i = 28$ from the treated sub-model $\mathcal{M}_{\text{tr}}$.

consistent with an immediate reduction in reported receipts. Specifically, in the first days the distribution is shifting up and below zero while, from $t = 32$ on stays below zero with very compact interquartile ranges (IQR), indicating a systematic reduction in receipts after audit. Occasional right-skewed distribution spikes suggest a very small subset of firms exhibiting temporary increases or lagged response. The presence of the zeros in the distribution is not completely unexpected in our setting as previously stated in reference with Figure 1.5. As previously stated there are just few post intervention days, discrete counts, and moderate overdispersion, making posterior uncertainty at the day level higher than desired. Overall, the shape and location of the distributions support a negative but heterogeneous short-run response. Finally, figure 1.7 provides an in-sample posterior predictive check for the cross-sectional mean. The posterior predictive means, from sub-models, track closely observed daily averages with the two nearly overlapping on most dates and the 95% predictive band containing the observations throughout. The shift at $t = 28$ is captured by transitioning from the control to the treated sub-model. Taken together, these results indicate that the treated/control sub-models are well calibrated for the first moment and accurately capture both the timing and magnitude of the shift.

1.7 Conclusions

This research work introduces a mean-based estimator for individual treatment effects (ITEs) in panel data settings. The proposed method address challenges in observational causal inference making the following key contributions: (i) relaxing the parallel trends assumption conditional on latent factors, (ii) capturing temporal dynamics, (iii) generalizing to all outcome distributions belonging to the exponential family, and (iv) handling staggered treatment adoption. Our methodological approach consist in fitting treated and control sub-models and then evaluating the ITE as a difference in posterior estimates of the models leading to lower uncertainty while maintaining coverage. Extensive MCMC simulations across Gaussian, Bernoulli, and Negative Binomial outcomes demonstrate improvements in RMSE and Mean Interval Scores relative to standard and commonly used estimator.

The benefits of the proposed approach is confirmed through an empirical application that focus on evaluating compliance behavior of firms working in the Greek energy sector post audit intervention. Through this application we further demonstrate the practical relevance of the method and the interpretability of the estimates obtained from the proposed estimator. From an applicative side-view, using data on approximately 200 firms, we show clear evidence of reduction of micro receipt emission in the post audit window, highlighting the importance of this intervention in preventing fraud. As a matter of fact, posterior estimates deliver tighter credible intervals than standard approaches, confirming an economically meaningful impact in fraudulent receipts reduction. This was also visible at raw data level where untreated firms maintained relatively dispersed receipt distributions, while treated firms exhibited notably compressed distributions concentrated at lower values after auditing.

However, our analysis is limited to a five-day post-audit window, which constrains inference about treatment persistence and whether the firms behaviour persist or attenuate over longer time. Despite these limitations, these findings make important contributions to the audit response literature providing a valuable method for policy effect estimates. The approach applicability across outcome distributions and treatment timing structures positions it as a practical tool for researchers and policymakers seeking credible, computationally efficient treatment effect estimates in observational panel data. Future work extending the post-audit observation window and examining treatment heterogeneity by firm characteristics could strengthen understanding of audit impact mechanisms and thus, inform more targeted compliance strategies.

In addition to these contribution, the works opens up to further methodolocgical and

empirical development. Future research could explore other functional form assumption in order to further relax the assumption and, moreover, instead of focusing on ITE estimation an estimator for heterogeneous causal effect may be considered or generally, further work could investigate systematic heterogeneity by firm size, sector, prior compliance history, or exposure intensity even using ITE, allowing policymakers to identify which types of firms respond most strongly to audits and to design more targeted enforcement strategies. From an empirical standpoint, the most important thing is to extend the post-audit period in order to monitor the treatment effect persistence in time. Such an extension would be particularly valuable for understanding whether audits generate lasting compliance improvements or short-term behavioural adjustments. Overall, these extensions would further enhance the practical relevance of the proposed estimator and improve its role as a versatile tool for causal inference in observational panel data environments.

Chapter 2

A graphical approach for the evaluation of modelling choices in epidemiological risk estimation

2.1 Introduction

In epidemiological research, as in statistical modelling, a substantial part of analytical results is intrinsically related to the choices made prior to model estimates (Baldwin *et al.*, 2022). Every modelling process includes a series of explicit or implicit decisions that shape how evidence is generated, summarized and interpreted. These decisions may include variables selection, the handling of missing data and outliers, as well as the specification of functional forms that describe relationships among variables. In general, the process of model building is not purely mechanical since it requires subjective judgment, theoretical reasoning, and familiarity with the subject matter (**?**). As discussed by Gelman and Loken (2013), even well intentioned and calibrated analytical decisions can lead to substantially different results, reflecting the wide number of reasonable paths that are available to a researcher. Consequently, researchers analyzing the same dataset may obtain different results simply because they made different, even if equally reasonable, modelling choices (Botvinik-Nezer *et al.*, 2020). This analytical flexibility has been recognized as one of the main factors contributing to the broader replication crisis across scientific disciplines (Errington *et al.*, 2021, Camerer *et al.*, 2016, Open Science Collaboration, 2015). Recent replication studies in psychology, economics, and clinical research have demonstrated that many published findings fail to replicate, raising concerns about the robustness of empirical evidence and the transparency of analytical practices (Micheloud and Held,

2024; Macrì Demartino *et al.*, 2024; Micheloud *et al.*, 2023; Pawel and Held, 2022; Held, 2020, among others). The problem of model specification has been analyzed from an applied perspective across various disciplines including epidemiology (Ioannidis, 2008), economics (Leamer, 1985), psychology (Bastiaansen *et al.*, 2020). In all these fields, statistical models are used by researchers to mathematically represent real-world phenomena, and, as previously stated, the resulting inferences depend critically on how these models are specified. This is primarily because no model can capture every aspect of reality. To simplify the overall process, researchers make assumptions about different aspects of the model based on limited data and knowledge. Recognizing the dependence of the results on analytical decisions and developing transparent frameworks that make such dependencies more explicit is therefore essential for understanding the sources of uncertainty in statistical modelling (Hoffmann *et al.*, 2021). In epidemiological context, different specifications have direct implications in the interpretation of estimated risks, influencing both the assessment of the risk and the reliability of subsequent predictions. This may have critical consequences since, the findings are subsequently used as a guide in prevention strategies or health policy decisions. A recent study conducted in nutritional epidemiology by Vorland *et al.* (2025) showed how even minor analytical choices, such as exposure definitions, variable configurations or covariates inclusion, can lead to substantial variations in estimates, highlighting how analytical flexibility can shape epidemiological conclusions and, if unaccounted for, may compromise the stability and reproducibility of evidence used in policy and guideline development.

To explore this problem, several methodological approaches have been proposed to systematically assess the impact of analytical decisions on research findings, including multiverse analysis (MA) (Del Giudice and Gangestad, 2021; Steegen *et al.*, 2016), specification curve analysis (SCA) (Simonsohn *et al.*, 2020), and the vibration of effects (VoE) framework (Patel *et al.*, 2015). These methods, provide a structured framework to investigate how analytical choices affects empirical findings by mapping the distribution of the outcome of the analysis, for example the estimated exposure-outcome association across a set of reasonable analytical choices. However, comparing different model specifications can be very challenging, as the number of possible combinations increases rapidly when multiple aspects are taken into account. Moreover, this complexity can lead to serious difficulties while attempting to compare epidemiological metrics across all scenarios and assess the overall consistency of the results. Inspired by the idea of MA, in this work, we propose a graphical approach for the systematic evaluation of modelling choices and their impact on results. The method provides insights into the robustness and stability of epidemiological conclusions while addressing the challenges associated

with multiple model comparisons.

The approach is illustrated through an empirical case study investigating the relationship between smoking and alcohol consumption, in terms of duration and intensity, and the risk of developing cancers of the upper aerodigestive tract (UADT), specifically those of the oral cavity, esophagus, and larynx. Tobacco smoking and alcohol consumption are well-established risk factors for UADT tumors (Bravi *et al.*, 2021), together accounting for approximately three quarters of all cases in developed countries, with evidence of a multiplicative effect when both exposures occur together (Pelucchi *et al.*, 2008; Polesel *et al.*, 2008). Correctly shaping the relationship between disease development and exposure factors in UADT cancer may be difficult as these tumors arise from multifactorial and interacting carcinogenic mechanisms that are not easily captured by simple models. In recent years, several studies have highlighted that the relationship between exposure and risk does not necessarily follow a linear path showing that the shape of the dose response function can vary across levels of exposure intensity and duration (Polesel *et al.*, 2005; Di Credico *et al.*, 2019, 2020, among others). Moreover, the assessment of duration and intensity is usually subjected to the presence of extreme values and biases. A common practice is avoid sparse data and outliers by limiting the analysis to some reasonable values for duration and intensity of exposure through percentile based cuts. In addition, a common occurrence is the tendency of participants to round their self-reported consumption to convenient values (i.e., multiples of 5 or 10), known as digit preference, that can introduce nonrandom measurement error, further affecting the estimation of dose–response relationships (Franceschi *et al.*, 1990). Taken together, these aspects underscore that the modelling of this epidemiological phenomena is subject to analytical choices that can have a notable impact on the risk estimates. Since these modelling decisions are affecting each others, their combined effect is difficult to measure. Thus, a systematic approach is needed to explore how alternative modelling specifications influence the results. In the proposed graphical framework the aim is to assess whether and how these choices are impacting the estimated risks for a set of prototype individuals representing different combinations of alcohol and tobacco use patterns, compared to a baseline individual with no exposure to either factor. In particular, the focus is on the nonlinerity of the relationship between exposures and cancer risk, modeled through univariate and bivariate splines in our formulation and compared to a classical linear effect formulation. Moreover, we also assess the impact of percentile based cuts during data cleaning while also examining how the inclusion or exclusion of extreme percentiles may introduce issues related to digit preference. As shown in the results, the approach clearly allows to assess how variations in model specification influence the robustness

and stability of the estimated risks. It provides a reliable graphical tool to evaluate model uncertainty and transparency that maps the variability of results across multiple plausible model configurations.

2.2 Multiverse analysis framework

Prior to the actual process of data analysis, decisions regarding model specification must be made. As extensively discussed in the previous section, there is not a single correct specification for a given research problem; instead, a wide range of reasonable and theoretically valid modelling options typically exist. Early approaches of fitting multiple models with different specifications emerged through sensitivity analysis (Leamer, 1985) and multimodel inference, contrasting the traditional practice of constraining the analysis to a single model. While specifing a single model, iterative stepwise procedures are used to optimize the fit of the chosen statistical model without considering alternative specifications; they can be particularly problematic, as there is no guarantee that the selected model is optimal and, more importantly it leads to ignore the uncertainty due to model choice. The MA methods have similar foundations as sensitivity analysis or multimodel inference and were originally proposed to systematically review and compare the results obtained from a wide range of statistically valid models in empirical research. The approach aims at identifying a set of plausible analytical specification, known as the multiverse, and subsequently evaluating an estimand across these specifications reporting all results using graphical visualization tools. The idea is similar to SCA and VoE, but it differs in its primary scope. While SCA focuses on whether an estimated effect persists across a range of plausible model specifications, and VoE emphasizes how the effect size varies, MA tries to understand what overall conclusions can be drawn across all defensible analytical choices. Generally speaking, the main idea behind this approaches is that if a reported effect is truly robust, it should remain stable across a wide range of analytically rational choices. In contrast, if results vary considerably depending on model specification, this variability itself becomes informative revealing which decisions materially influence the conclusions. Once the analysis are performed across the set of plausible models conclusions may be drawn differently; for example, one may focus on evaluating the percentage of significant results (Young, 2018), assess variations in estimated effects qualitatively (Maie *et al.*, 2024), or examine them quantitatively or graphically (Constantino *et al.*, 2021). In this work, we leverage MA through graphical tools to visually assess how results vary across reasonable alternative analytical decisions and how these variations relate to specific modelling choices.

Let $\mathcal{D} = \{Y_i, E_i, X_i\}_{i=1}^{n}$ denote a dataset containing the outcome $\mathbf{Y}$, the exposures of interest $\mathbf{E}$, and a set of covariates $\mathbf{X}$. Let $\theta_{i,s} \in \Theta$ denote the estimand of interest for i-th unit, where $s \in \mathcal{S}$ index a specific model definition belonging to the specification set space $\mathcal{S}$. The specification set space refers to all considered and statistically valid analytical decisions. In our specifics case, these decisions include data preprocessing rules $\mathcal{T}$ and choices regarding the functional form describing the relationship between $\mathbf{Y}$ and $\mathbf{E}$, denoted as $\mathcal{F}$. Other specifications, such as covariate selection may be also included. The full specification set space $\mathcal{S}$ is formally defined as the collection of all combinations of the chosen specifications:

$$\mathcal{S} = \mathcal{F} \times \mathcal{T}$$

For each specification $s \in \mathcal{S}$ a specific statistical model $\mathcal{M}_s$ is fitted and subsequently the parameter of interest is estimated $\hat{\theta}_s$ along with its respective standard error $\widehat{\text{se}}(\hat{\theta}_s)$. Each model $\mathcal{M}_s$ can be referred to as a multiversal model, representing one possible analytical universe among all reasonable alternatives. Practically, MA involves mapping each analytical choice to an empirical estimate $s \longmapsto \hat{\theta}_s$. Rather than committing to a single specification a priori, estimates across all defensible specifications are provided, subsequently the full collection $\{s, \hat{\theta}_s\}$ is examined; this collection represents the final output of the MA. The results are graphically summarized plotting the parameter of interest $\hat{\theta}_s$ along one axis, typically ordered by magnitude, with each point corresponding to the specific analytical decisions that generated it. These specification choices are reported below the curve, enabling readers to understand which analytical choices produced each estimates. In this way the representation reveals the extent to which the results depend on model specification, identifies which findings are robust across analytical choices, and demonstrates which conclusions are sensitive to specification decisions.

A key characteristic that makes MA differ from standard model comparison techniques is that the analytical choices made are not independent of one another. When deciding whether to control for a variable, this decision may be influenced by prior decisions about other variables in the model. This creates a system of conditional relationships among specifications rather than a collection of independently chosen alternatives. This has important implications since the statistical properties assumed for independent model selection cannot be directly applied. Instead, the multiverse represents an interconnected network of decisions, where changing one specification automatically affect other similar specifications.

The method becomes particularly valuable when researchers face multiple plausible ways to model the data, the study involves observational rather than experimental data (requiring many subjective decisions about variables and transformations), the research findings will influence important decisions or policies, or analytical flexibility is substantial (as in studies with numerous potential confounders). By contrast, multiverse analysis may be less useful when a study design is highly structured with few defensible options, strong theoretical consensus guides model selection, or computational constraints are severe. In fields like epidemiology and social science, where analytical decisions are numerous and consequential, MA plays an important role in documenting and displaying the impact of these inevitable choices on research conclusions. Before generating the multiverse, researchers must establish clear criteria for which analytical options qualify as reasonable. A valid and statistically defensible specification should meet several requirements; (i) each choice should rest on theoretical or empirical justification so, it must not be selected just because it produces the desired results, (ii) the specification must be statistically appropriate, satisfying model assumptions and producing interpretable estimates and (iii) redundant or overlapping specifications should be eliminated to maintain a parsimonious multiverse. The main steps of a MA are summarized below:

1. **Universe space definition:** Define all defensible analytical choices before analyzing the data. This includes: data preprocessing rules, functional form choices, covariate selection strategies and other domain specific analytical decisions along with the definition of the parameter of interest, θ.

2. **Universe generation and analysis:** Execute all valid combinations of analytical specifications across the specification space $\mathcal{S}$. For each universe $s \in \mathcal{S}$, fit the statistical model $\mathcal{M}_s$ and estimate the parameter of interest $\hat{\theta}_s$ along with its standard error $\widehat{\text{se}}(\hat{\theta}_s)$.

3. **Outcome Sensitivity Assessment:** Examine the distribution of the outcome of interest θ across all universes. This enables to assess which results are more sensitive to on analytical choices and whether findings are robust to specification decisions.

4. **Model-Outcome Connection:** Identify which analytical choices drive outcome sensitivity by assessing the stability of the analytical choices.

5. **Multiverse validation:** Critically evaluate whether all included universes remain equally defensible using an adequate validity metrics (i.e. model fit statistics) in

order to identify whether some specifications produce systematically poor-fitting models. This has received limited attention in the literature.

6. **Interpretation and reporting**: Summarize findings and communicate results through appropriate graphical and non graphical tools.

The approach does not attempt at identifying the best model, instead it allows to reveal how sensitive are the results with respect to analytical choices. It was fundamentally designed as a transparency tool rather than a model selection procedure. Anyway, in some cases, the researcher may require to make a final choice obtaining a single estimate for the parameter of interest that will be used for policy, publication, or decision-making processes. In this cases, rather than select a single best model MA can employ model averaging approaches, such as Bayesian model averaging, to synthesize results across multiple specifications; so, when needed an extra seventh step can be added to the previous listed steps. It is unclear, however, under which circumstances model averaging improve on using a single model (Dormann *et al.*, 2018).

An important challenge in implementing MA is that the set of analytically plausible specifications is typically quite large, sometimes numbering in the hundreds or thousands. This creates a fundamental problem since it becomes difficult to visualize and interpret each graphical representation as the number of considered models increase. Consequently, researcher must aim at balancing specification comprehensiveness and reporting clarity. Interactive visualization tools or tailored graphical strategies can be particularly useful in these cases. Despite these practical constraints, MA is a valuable method that substitutes the illusion of a single correct model with transparent acknowledgment of analytical flexibility and its impact on empirical conclusions. A useful perspective is to view the multiverse as a discrete approximation to a richer modelling space. In semiparametric and nonparametric statistics, uncertainty is often represented by allowing parts of the model (e.g., smooth functions) to vary in infinite dimensional spaces; multiverse analysis operationalizes a related idea by evaluating a finite, defensible subset of such possibilities.

2.3 Data and model choice

In this section we describe the main analytical decisions that have been considered, and therefore the resulting set of model specifications included in the analysis. These decisions are taken considering all kind of choices that may appear particularly relevant for our empirical case. In fact, as mentioned previously, we focus on two broad classes of

decisions that may have a significant impact on the result. In particular, we consider: (i) data preprocessing decisions related to percentile based trimming of the exposure variables, and (ii) modelling choices concerning the use of non-linear functional forms to represent the association between exposures and the risk of UATD cancer. The full set of specifications is obtained by combining these decisions across all possible configurations, thereby reflecting the multiplicity of plausible analytical choices. In addition to evaluating all possible combinations of trimming rules and functional forms, we also analyze a scenario in which non linear effect models are fitted without applying any percentile cut. This allows us to highlight the effect of the non linear functional form alone and to compare results obtained from the full dataset. The analyses were conducted by jointly considering preprocessing and modelling choices that will be described the in following sections, resulting in a multiverse of 1,944 model specifications.

2.3.1 Data preprocessing

The main exposures for the empirical study are respectively smoking and alcohol consumption, each measured along two dimensions: intensity (i.e. number of cigarettes per day and milliliter of ethanol per day) and duration (i.e. years of consumption). Both variables typically exhibit skewed distributions, with heavy upper tails arising from a small number of individuals reporting extremely high levels of exposure. These extreme values may impact significantly on model estimates and subsequent interpretation. Generally, to avoid this issues, percentile based cuts are applied on exposure variables to make the data distribution less skew. The choice of the percentile level is also a choice that is generally made upon some personal beliefs of the researchers, generally proposing cuts on 95% or 99%. In our case we are going to take into consideration both levels while also considering the "no-cut" option. All combinations of cuts among the four exposure variables are considered. A second consideration concerns potential digit preference in self-reported intensity values, where participants tend to round consumption to convenient units. To assess the influence of this phenomenon on model estimates, we treat the inclusion or exclusion of rounded extreme values when they overlap with a specific percentile, as an additional analytical choice.

2.3.2 Model specification

Beyond preprocessing decisions, which may be particularly relevant in epidemiology, another concern is how exposures are represented in the regression model. In this specific case we are interested in understanding the importance of considering the

potential nonlinearity of the dose response (exposure-risk) relationship. When exposures are continuous, standard parametric models may fail to capture complex or nonlinear exposure–response relationships. Flexible approaches based on regression splines have therefore been widely adopted in epidemiological research (Rosenberg *et al.*, 2003). For this aim we rely on Generalized Additive Models (GAM) (Hastie and Tibshirani, 1986) which can incorporate both linear and non linear terms. In particular, to capture non-linearity we use spline terms, which enable to model flexible exposure response relationships without imposing a specific functional form. Technically, splines are multiple polynomial functions joined smoothly at predefined locations called knots. Within each interval between knots, the function is described by a low-degree polynomial, while smoothness constraints ensure that these pieces connect seamlessly. This construction allows splines to adapt to local changes in the shape of the relationship, providing enough flexibility to capture complex patterns while avoiding the instability associated with using high degree global polynomials. Splines are therefore particularly suited to epidemiological modelling, where the true underlying relationship is often unknown and unlikely to follow a simple linear trend. We considered two type of spline representation: (i) univariate splines that are applied separately to intensity and durations for both alcohol and smoke consumption and, (ii) bivariate splines where we jointly model the effect of intensity and duration, separately for smoking and for alcohol.

2.4 Empirical application

In this section, cancer risk, in terms of odds ratio (OR), is estimated using results from regression based epidemiological models that associate smoking and alcohol consumption to the probability of developing UADT cancers. The outcome of interest is cancer status, modeled as a binary response. Exposures include smoking intensity and duration, as well as alcohol consumption intensity and duration. Demographic variables, specifically age, education, and sex, are included as control covariates, with age and education modeled using smooth functions. GAMs are employed to capture nonlinear exposure–response relationships using penalized regression splines while, bivariate relationships are modeled using tensor product interaction smooths. All smooth terms are specified accordingly to the R package `mgcv` (Wood, 2001). Model estimation is conducted across all datasets obtained by jointly varying exposure cut-offs and inferential specifications, resulting in a multiverse of 1,944 model configurations (see Section B.1 for details).

2.4.1 Data source

The data were obtained from a series of hospital-based case-control studies on cancer of the UADT conducted between June 1986 and June 1989 in three Italian regions: Pordenone (north-eastern Italy), Latina (central Italy), and the greater Milan area in Lombardy. Cases included men and women aged 82 years or younger with confirmed diagnoses of oral cavity, laryngeal, or esophageal cancer. Control participants were patients younger than 80 years admitted to the same hospitals for non-malignant conditions. Data on smoking and alcohol consumption habits, along with sociodemographic and clinical characteristics, were collected through structured interviews. The sample included 1710 cancer cases and 4323 control cases. In some cases, the same control unit has been used on more than one cancer site. Detailed distribution of cases and controls stratified by cancer site, age, sociodemographic characteristic and exposure patterns are shown in Table 2.1. The final analytical sample included participants with complete information on smoking and alcohol consumption and relevant sociodemographic variables. Subjects with implausible exposure values or missing data on key variables were excluded according to systematic criteria (see Figure 2.1).

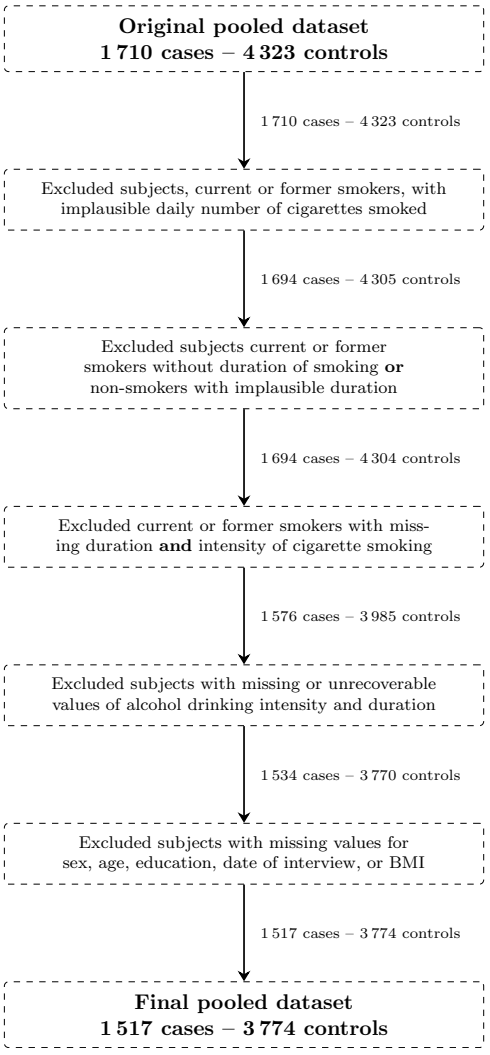


FIGURE 2.1: Flowchart of study population selection. The diagram shows the sequential exclusion criteria applied to the original pooled dataset, resulting in the final analytical sample of 1,517 cases and 3,744 controls.

TABLE 2.1: Distribution of 3,744 controls and 1,517 cancer cases by cancer site according to sex, age, education, center, tobacco smoking and alcohol drinking habits.

	Oral cavity				Larynx				Esophagus			
	Controls		Cases		Controls		Cases		Controls		Cases	
	n	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)
Age (years)												
< 40	162	7.3	37	4.3	4	0.9	1	0.6	12	1.2	6	1.4
40–44	123	5.5	37	4.3	13	2.9	6	3.4	22	2.1	9	2.1
45–49	257	12.0	99	11	32	7.1	16	8.9	77	7.4	36	8.3
50–54	324	15.0	128	15	76	17.0	35	20.0	130	12.0	52	12.0
55–59	331	15.0	160	18	104	23.0	41	23.0	204	20.0	82	19.0
60–64	358	16.0	152	17	88	20.0	38	21.0	225	22.0	94	22.0
65–69	377	17.0	150	17	79	18.0	26	15.0	210	20.0	86	20.0
70–74	257	12.0	89	10	52	12.0	16	8.9	154	15.0	63	15.0
≥ 75	37	1.7	17	2.0	0	0.0	0	0.0	9	0.9	5	1.2
Sex												
Male	1356	61.0	693	80	317	71.0	156	87.0	825	79.0	391	90.0
Female	870	39.0	176	20	131	29.0	23	13.0	218	21.0	42	9.7
Education												
No education	15	0.7	8	0.9	5	1.1	2	1.1	0	0.0	2	0.5
Elementary school	1095	49.0	481	55	254	57.0	105	59.0	592	57.0	248	57.0
≤Junior high school	563	25.0	230	26	119	27.0	45	25.0	269	26.0	105	24.0
High school graduate	431	19.0	115	13	46	10.0	22	12.0	137	13.0	56	13.0
Technical sch., some college	26	1.2	5	0.6	6	1.3	1	0.6	12	1.2	3	0.7
College graduate	96	4.3	30	3.5	18	4.0	4	2.2	33	3.2	19	4.4
Center												
NE	1037	47.0	467	54	334	75.0	141	79.0	904	87.0	394	91.0
MI	762	34.0	300	35	114	25.0	38	21.0	139	13.0	39	9.0
LT	427	19.0	102	12	0	0.0	0	0.0	0	0.0	0	0.0
Smoking status												
Never	969	44.0	124	14	243	54.0	32	18.0	394	38.0	18	4.2
Former	671	30.0	231	27	0	0.0	0	0.0	375	36.0	143	33.0
Current	586	26.0	514	59	205	46.0	147	82.0	274	26.0	272	63.0
Smoking intensity (cig./day)												
1–10	450	36.0	122	16	68	33.0	27	18.0	238	37.0	56	13.0
10–20	564	45.0	394	53	101	49.0	82	56.0	282	43.0	237	57.0
20–30	116	9.2	134	18	25	12.0	24	16.0	79	12.0	87	21.0
30–40	87	6.9	74	9.9	11	5.4	14	9.5	40	6.2	27	6.5
> 40	40	3.2	21	2.8	0	0.0	0	0.0	10	1.5	8	1.9
Smoking duration (years)												
1–10	106	8.4	13	1.7	2	1.0	0	0.0	45	6.9	4	1.0
11–20	242	19.0	51	6.8	14	6.8	1	0.7	105	16.0	13	3.1
21–30	336	27.0	156	21.0	35	17.0	18	12.0	143	22.0	74	18.0
31–40	295	23.0	280	38.0	69	34.0	66	45.0	183	28.0	134	32.0
> 40	278	22.0	245	33.0	85	41.0	62	42.0	173	27.0	190	46.0
Drinking status												
Never	430	19.0	65	7.5	47	10.0	4	2.2	77	7.4	17	3.9
Former	118	5.3	139	16.0	24	5.4	24	13.0	73	7.0	51	12.0
Current	1678	75.0	665	77.0	377	84.0	151	84.0	893	86.0	365	84.0
Drinking intensity (drinks/day)												
< 5	1260	70.0	298	37.0	230	57.0	30	17.0	568	59.0	99	24.0
5–10	471	26.0	284	35.0	142	35.0	98	56.0	346	36.0	228	55.0
11–15	46	2.6	151	19.0	19	4.7	28	16.0	43	4.5	71	17.0
≥ 15	19	1.1	71	8.8	10	2.5	19	11.0	9	0.9	18	4.3
Drinking duration (years)												
1–10	55	3.1	12	1.5	3	0.7	0	0.0	11	1.1	6	1.4
11–20	139	7.7	52	6.5	12	3.0	8	4.6	43	4.5	25	6.0
21–30	384	21.0	162	20.0	70	17.0	29	17.0	135	14.0	61	15.0
31–40	520	29.0	254	32.0	136	34.0	71	41.0	310	32.0	126	30.0
> 40	698	39.0	324	40.0	180	45.0	67	38.0	467	48.0	198	48.0

2.4.2 Prototype subjects and risk assessment

To evaluate the estimated cancer risk across different exposure scenarios, we defined a set of prototype individuals representing distinct combinations of smoking and alcohol consumption patterns. These prototypes allow us to move beyond aggregate effect estimates and assess how model specification choices affect risk predictions for specific exposure profiles. Each prototype is characterized by four exposure dimensions: tobacco smoking intensity (cigarettes per day), smoking duration (years), alcohol intensity (ml ethanol per day), and alcohol duration (years), along with current, former, or never-user status for each substance. We use the following abbreviations for exposure variables: I_{sig} and D_{sig} denote the intensity and duration of cigarette smoking, respectively; I_{alc} and D_{alc} represent the intensity and duration of alcohol consumption; and S_{sig} S_{alc} indicate the exposure status (current, former, or never) for smoking and alcohol use, respectively. Table 2.2 presents the ten prototype profiles considered in this analysis. These profiles cover a range of realistic exposure scenarios, from individuals with no exposure to those with heavy, long-duration use of both tobacco and alcohol. The prototypes include patterns such as current smokers and drinkers at varying intensity levels, former users, and mixed exposure histories. For each prototype, we estimated the predicted risk of developing UADT cancer relative to the baseline individual (no smoking, no alcohol consumption) across all model specifications in the multiverse. This approach allows us to assess not only whether overall effect estimates are robust, but also whether conclusions about risk magnitude and patterns differ meaningfully depending on modelling choices. The selection of these specific prototypes was guided by: (i) ensuring realistic combinations based on observed patterns in the data, (ii) representing clinically relevant exposure scenarios, and (iii) cover a range from unexposed to high exposed individuals to capture heterogeneity in dose-response relationships. The risk for each model specification in the multiverse is estimated in terms of odds ratio (OR) representing the risk of UADT cancer for each prototype individual relative to a baseline individual with no exposure to either smoking or alcohol. The OR quantifies the multiplicative change in odds of disease for a given exposure profile compared to the reference category. We evaluated risk using 95% confidence intervals (CIs) computed from the standard errors of the estimated parameters. An OR estimate is considered statistically significant if the 95% CI does not include the null value (OR = 1).

TABLE 2.2: Prototype individuals used for cancer risk assessment across multiverse specifications. Profiles represent different combinations of smoking and alcohol consumption intensity, duration, and status.

Profile	I_{sig} (cigs/day)	D_{sig} (years)	I_{alc} (ml ethanol/day)	D_{alc} (years)	S_{sig}	S_{alc}
1 (Baseline)	0	0	0	0	Never	Never
2	5	20	15	35	Current	Current
3	5	40	15	35	Current	Current
4	15	20	15	25	Current	Current
5	15	50	15	35	Current	Current
6	25	40	15	35	Current	Current
7	35	40	20	45	Current	Current
8	5	10	20	35	Current	Former
9	15	20	15	35	Former	Current
10	15	20	15	35	Former	Never

2.4.3 Results

In Figures 2.2, 2.3, and 2.4, we report the results from the multiverse analysis for all three cancer sites obtained considering the different modelling scenarios described in Section 2.3.2. Each figure consists of three horizontally arranged panels sharing the same horizontal axis. The x-axis indexes the different model specifications included in the multiverse, where each number corresponds to a specific combination of functional-form choices (linear or spline) for the exposure variables; models are ordered according to their ΔAIC values. The top panel, divided into three rows, shows the estimated odds ratios and 95% confidence intervals for the prototype individuals described in Table 2.2 according to the model specifications with $\Delta\text{AIC} < 0$. Each point–interval therefore represents the OR estimate obtained under a given modelling choice, allowing the reader to visually assess the stability of the estimates across specifications. The middle panel shows a heatmap summarizing the modelling choices applied to each exposure variable for every specification, with darker cells indicating the use of spline terms and lighter cells indicating linear terms. The bottom panel displays the ΔAIC values comparing each proposed specification from the multiverse with the baseline linear model, where for the i-th specification $\Delta\text{AIC}_i = \text{AIC}_i - \text{AIC}_{\text{baseline}}$. Negative ΔAIC values indicate an improvement in model fit relative to the linear specification, while positive values indicate worse fit; since AIC penalizes model complexity, negative values imply that the increased flexibility of the spline specification is supported by the data. Across all cancer

sites, OR estimates for prototype individuals show a high degree of stability, suggesting that the estimated associations between smoking/alcohol exposure and UADT cancer risk are robust to the modelling choices considered.

For oral cavity cancer (Figure 2.2), the ΔAIC values indicate modest improvements in model fit when applying splines to alcohol intensity, smoking intensity, and smoking duration, especially in the best-performing models. Alcohol intensity is always nonlinear in the best models, suggesting that the exposure/ response relationship for this variable is not adequately described by a linear function. Despite the modelling differences, the estimated ORs remains stable as shown in upper panel, showing that all individuals experience moderate to high risks coherent with their habits with limited variations across models. Prototype subject 6, representing the worst habits in terms of alcohol and tobacco consumption, shows wider confidence intervals, reflecting data sparsity at extreme levels. No exposure profile shows a reversal of effect direction; current smokers and drinkers consistently show elevated risks, while former users show clearly lower risks.

For laryngeal cancer (Figure 2.3), model fit is more sensitive to specification choices. In particular, spline terms for alcohol intensity yield the largest reductions in AIC, producing what we refer to as a "ΔAIC jump." Removing the spline on alcohol intensity brings the ΔAIC values close to zero, indicating little advantage over the linear model. Nevertheless, OR estimates remain broadly consistent across specifications, even though risks are generally higher and confidence intervals wider than in the oral cavity site, reflecting both a stronger role of smoking and alcohol in the carcinogenic process and the smaller number of cases for this site. Occasional shifts in the relative ordering of prototype profiles do not alter the overall interpretation. For this site no former smokers where available in the dataset.

For esophageal cancer (Figure 2.4), modelling choices have the strongest influence on AIC. Spline terms for smoking and alcohol intensity substantially improve model fit, especially in the best specifications, while spline terms for durations play a minor role. This indicates that intensity of exposure is more critical than duration in capturing risk patterns for this site. OR estimates remain generally stable, with greater uncertainty at extreme exposure levels and somewhat more dispersed estimates for the highest-intensity subjects. Former smokers and drinkers consistently show very low ORs, and the qualitative ranking of profiles is comparable to the other cancer sites.

To evaluate the robustness of risk estimates with respect to the preprocessing choices described in Section 2.3.1, we generated additional heatmaps reporting ORs for all prototype individuals across all model specifications (Appendix B.2 for graphs and further explanation of how to read it). These confirm the stability of the risk estimates

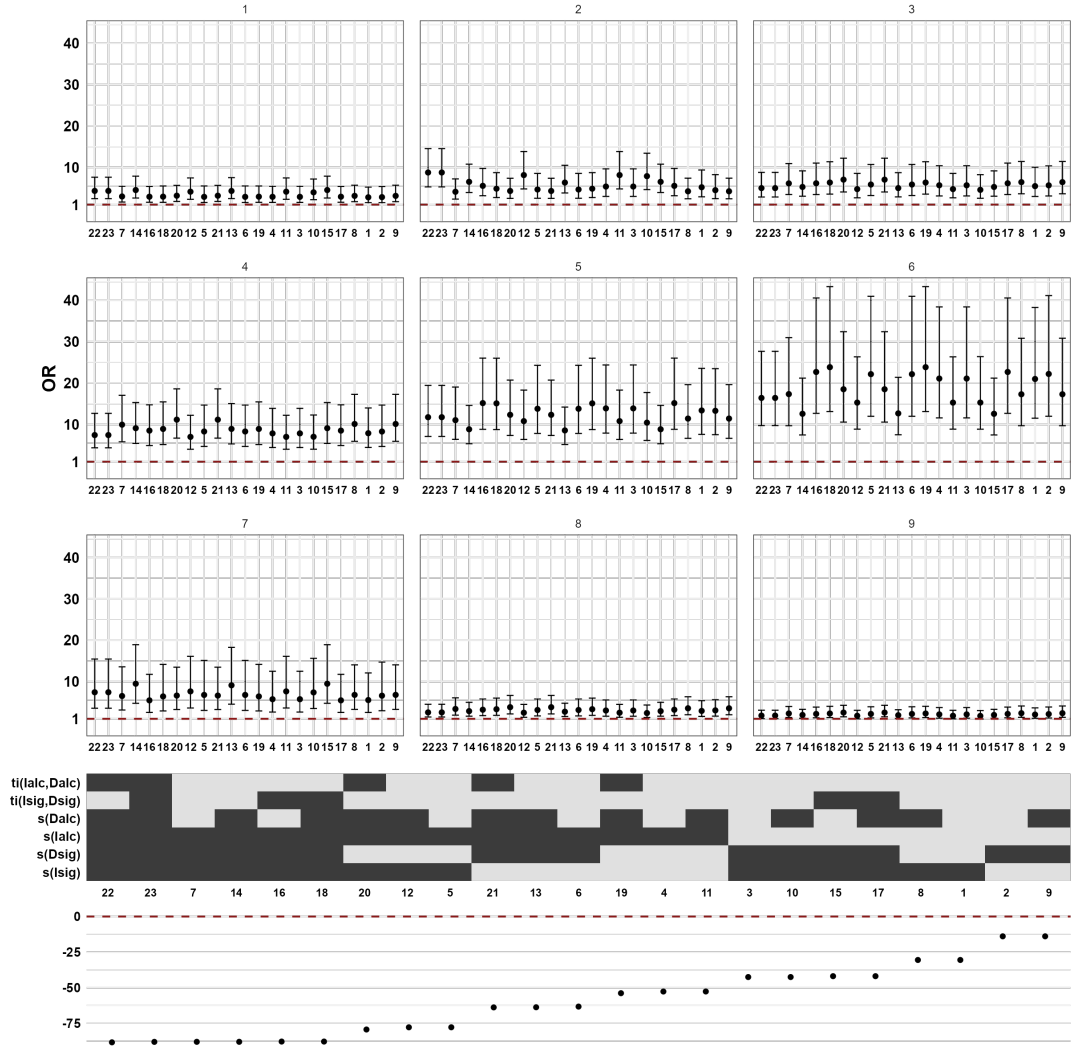


FIGURE 2.2: Multiverse assessment of model fit, specification choices and risk estimates for oral cavity cancer site. First panel shows OR estimates across considered models from the multiverse. Middle panel shows the combinations of spline terms (in black) active and not active (in grey). Bottom panel shows the ΔAIC of considered model and the baseline model with all linear terms.

and highlight an important methodological regarding the treatment of digit preference in our case for smoking intensity reporting. Prototype subject 6 shows noticeable variation in OR estimates when alternative preprocessing decisions are applied to the upper tail of smoking intensity values. Because many heavy smokers report rounded values, excluding observations that coincide with the percentile cut-offs (in our case, the 95th percentile corresponds to 40 cigarettes per day that are exactly two packages) would artificially reduce information on extreme exposures, strongly affecting risk estimates for heavy smokers. In this case risk estimates would be higher. This can be visualized as a clearly outlined pattern in panel a of Figure 2.5 that disappeared when these values are included (panel b).

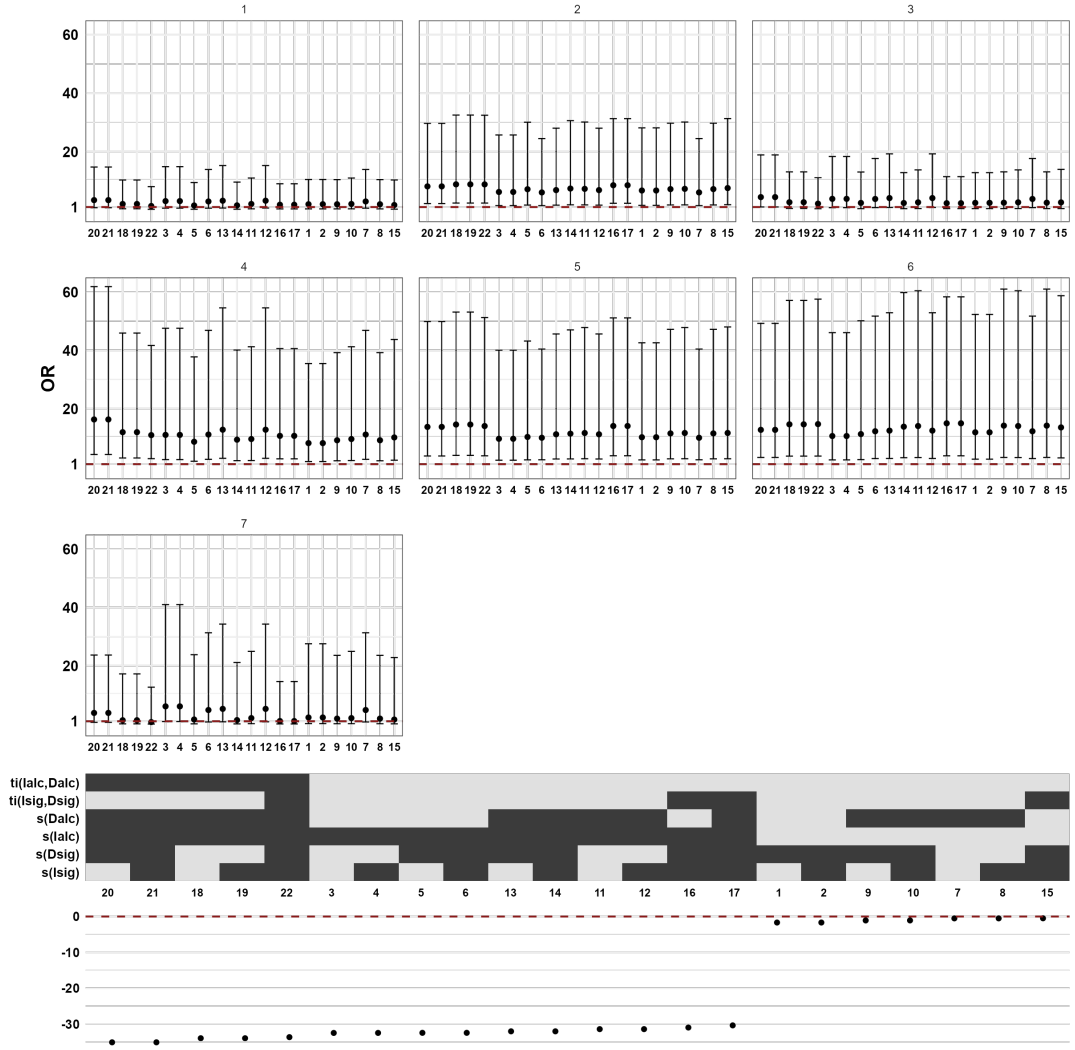


FIGURE 2.3: Multiverse assessment of model fit, specification choices and risk estimates for laryngeal cancer site. First panel shows OR estimates across considered models from the multiverse. Middle panel shows the combinations of spline terms (in black) active and not active (in grey). Bottom panel shows the ΔAIC of considered model and the baseline model with all linear terms.

Following the outcome sensitivity assessment we examined the stability of the analytical choices considered in the multiverse analysis. Spline based modelling of exposure intensity yielded substantial improvements in model fit when applied to alcohol consumption, particularly for laryngeal and esophageal cancers. As shown in Figures 2.6, 2.7, and 2.8, the estimated exposure-response curves reveal clear nonlinear relationships that would be missed under a standard linear specification, confirming insights from the previous analysis. For oral cavity cancer, smoking intensity and duration also exhibit nonlinearity. For smoking duration this is particularly pronounced when cutoff is absent or at the 99th percentile threshold cut. For the remaining exposures, preprocessing cutoffs have limited influence, and the functional form remains relatively stable across

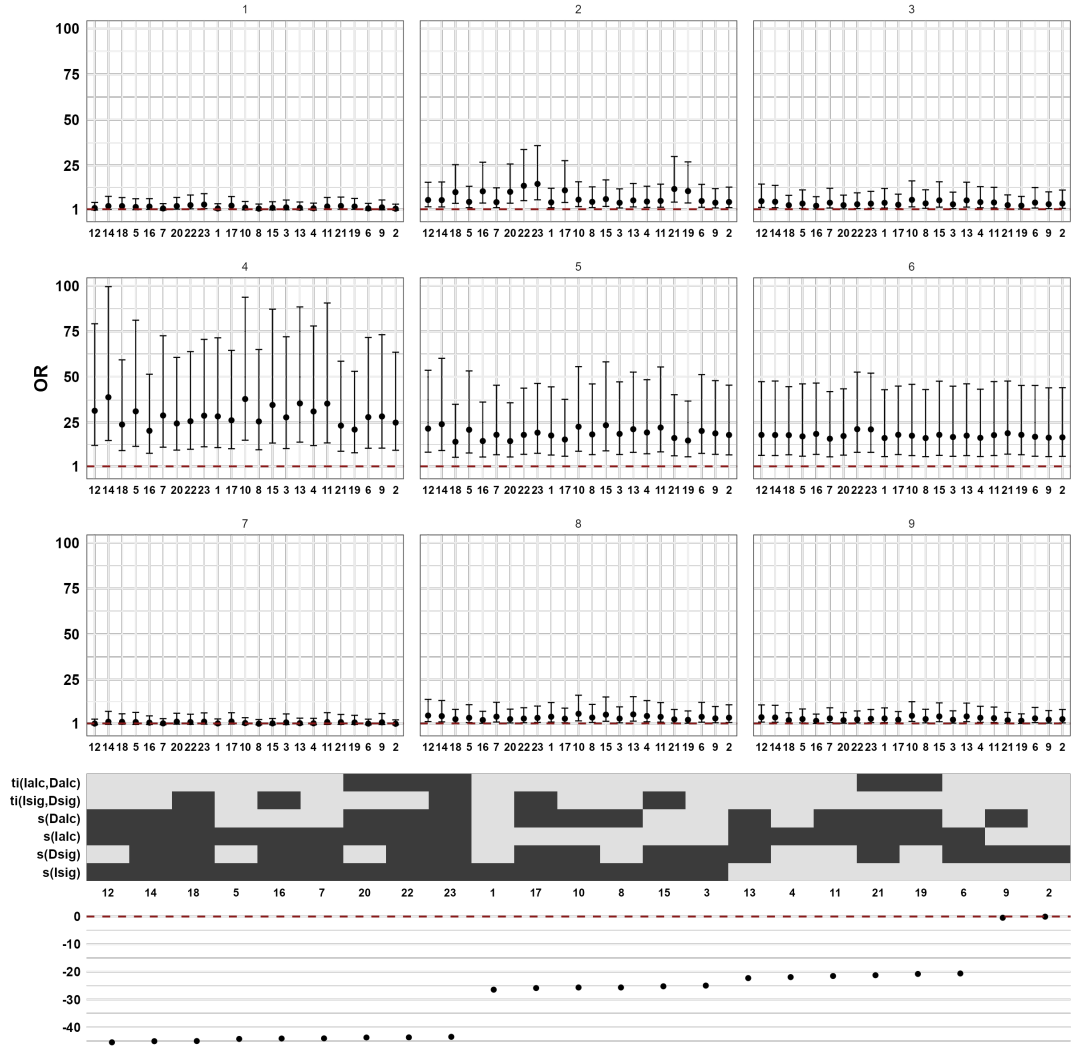
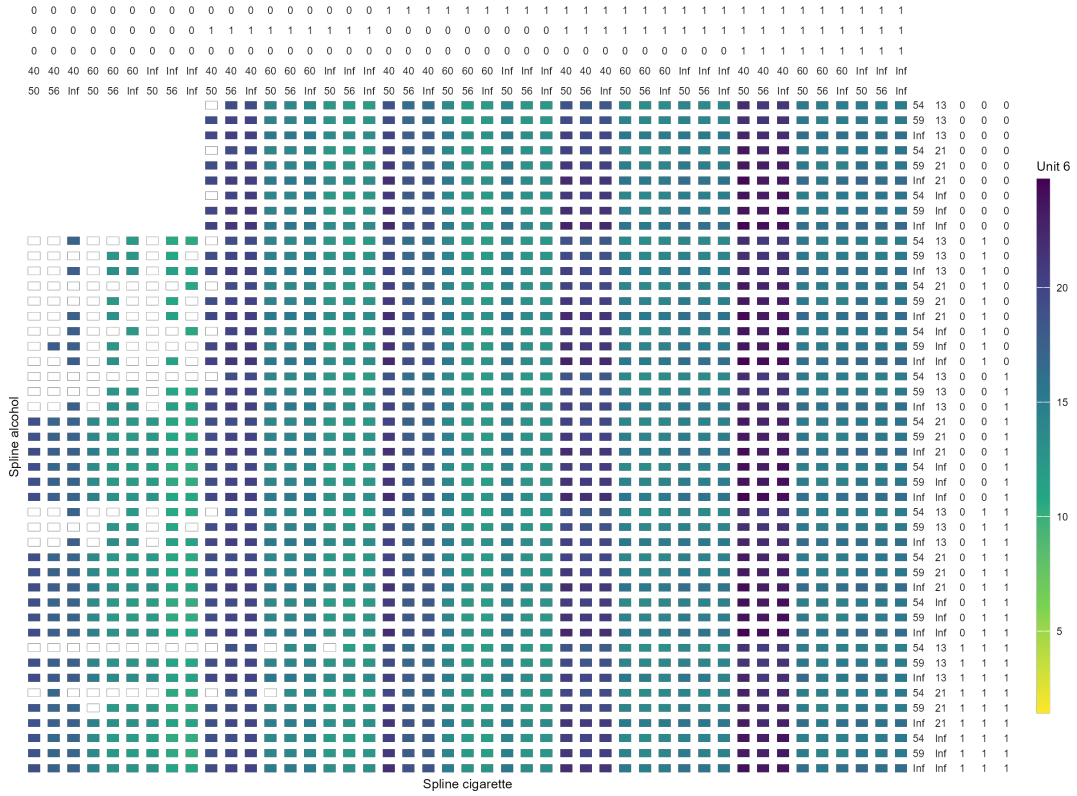


FIGURE 2.4: Multiverse assessment of model fit, specification choices and risk estimates for esophageal cancer site. First panel shows OR estimates across considered models from the multiverse. Middle panel shows the combinations of spline terms (in black) active and not active (in grey). Bottom panel shows the ΔAIC of considered model and the baseline model with all linear terms.

specifications. For laryngeal cancer, a pronounced nonlinearity for both smoking and alcohol exposure is observed. The spline curves reveal complex dose-response patterns that differ substantially from what linear models would predict, underscoring the importance of flexible functional form specification for accurate risk characterization corresponding also to the marked ΔAIC improvement previously noted. Moreover, alcohol intensity demonstrates a particular sensitivity to preprocessing decisions, with major differences in the linear predictor across cutoff strategies. Lastly, for esophageal cancer, spline modelling proves particularly relevant for alcohol consumption exposure and slightly for smoking duration. The functional form seems to remain robust across preprocessing decisions. Overall, this consistency across all three cancer sites indicates that conclusions



(a)

(b)

FIGURE 2.5: Comparison of OR estimates for prototype individual 6 across all multiverse specifications, distinguishing between exclusion (a) and inclusion (b) of percentile values during preprocessing.

about where flexible modelling is necessary are methodologically valid and do not strictly depend from data preprocessing choice and distribution. However, preprocessing specifications can still slightly influence linear predictor magnitudes in some cases reflecting the importance of a comprehensive analysis on both dimensions for robust conclusions.

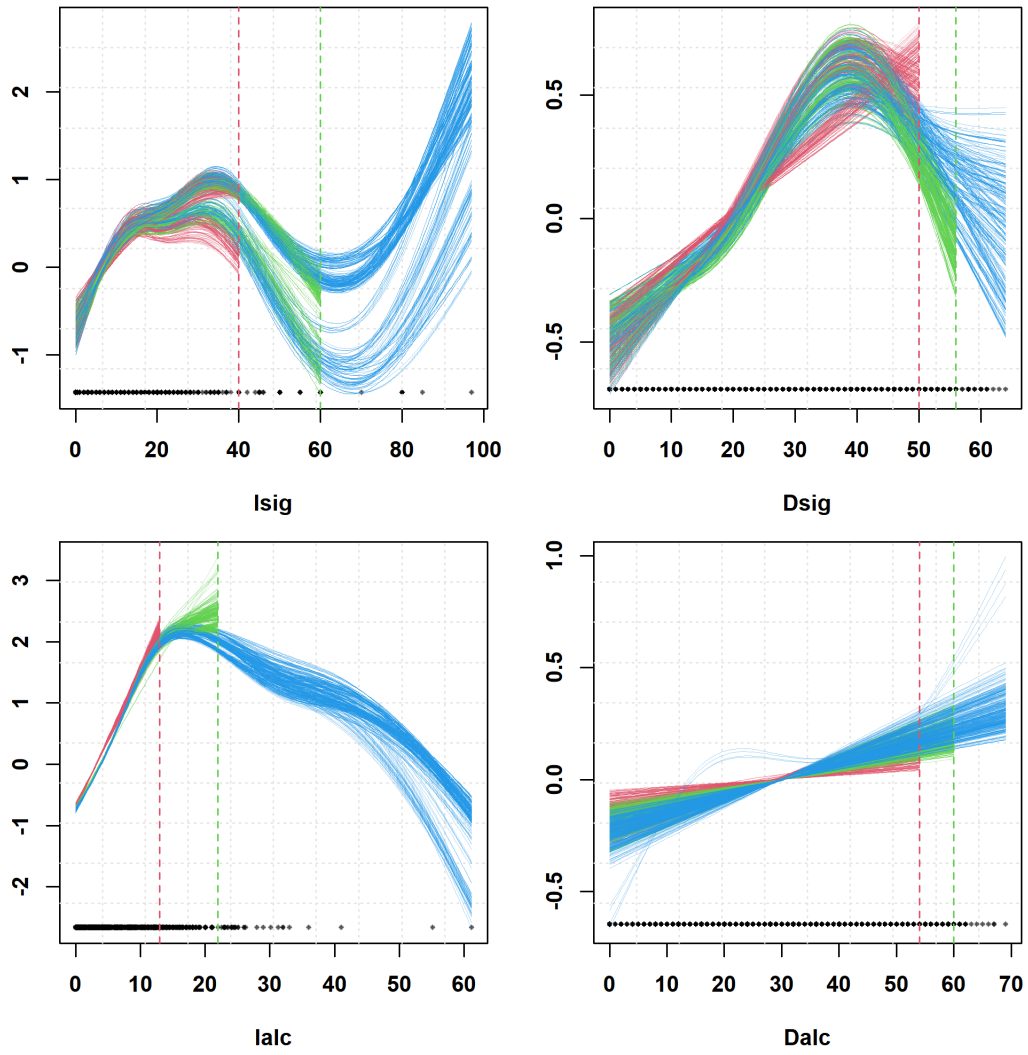


FIGURE 2.6: Stability of univariate spline based exposure–response curves across preprocessing and analytical choices for the oral cavity cancer site. Red lines correspond to models using a 95th-percentile cutoff on the exposure variable, green to a 99th-percentile cutoff, and blue to no cutoff.

To further examine the joint effects of smoking and alcohol exposures and assess the robustness of this two dimensional relationship, we generated contour plots of the bivariate spline terms for each cancer site. As shown in Figures 2.9–2.11, these plots display bivariate spline contour estimates for the joint effect of exposure intensity and duration on the model's linear predictor (link scale), allowing visualization of potential interaction patterns. Colors in the contour plots encode the magnitude of the bivariate

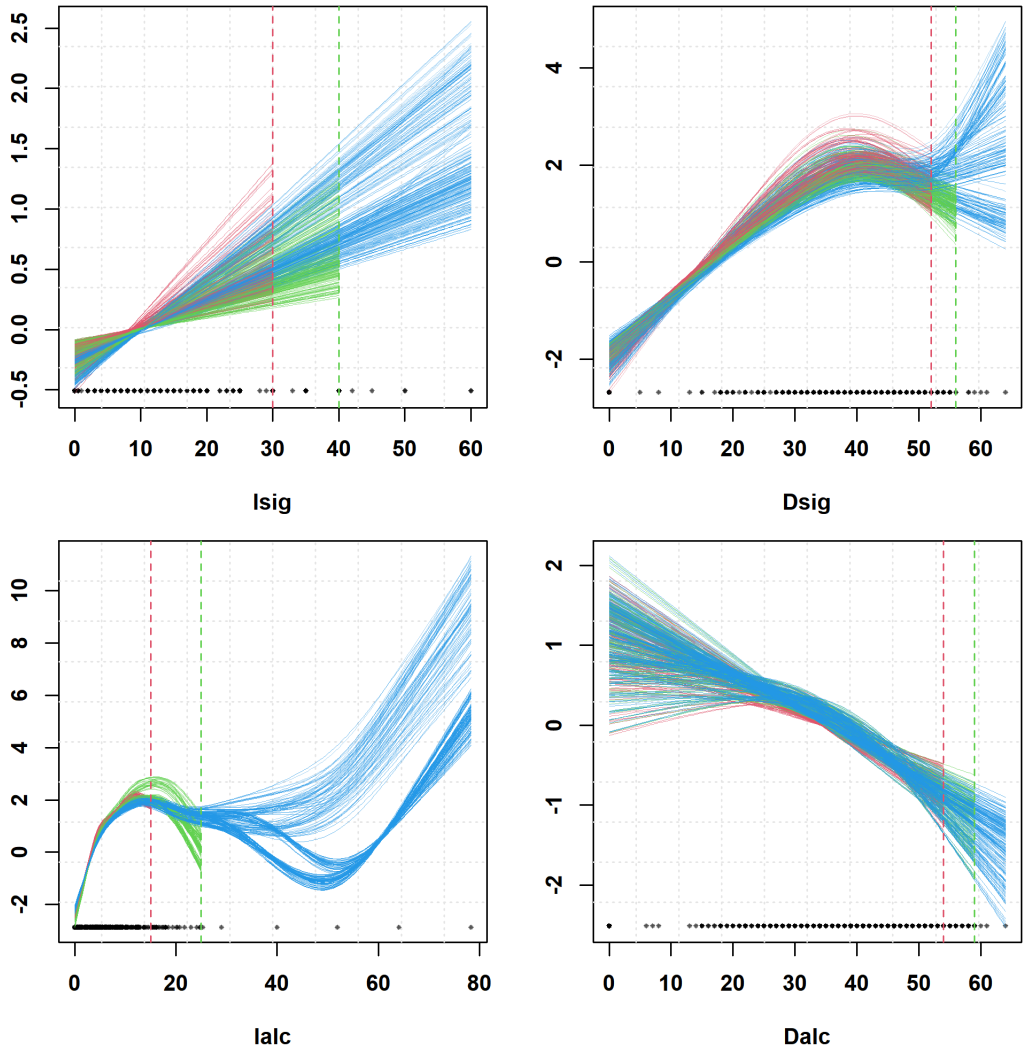


FIGURE 2.7: Stability of univariate spline based exposure–response curves across preprocessing and analytical choices for the laryngeal cancer site. Red lines correspond to models using a 95th-percentile cutoff on the exposure variable, green to a 99th-percentile cutoff, and blue to no cutoff.

spline contribution to the linear predictor, rather than predicted probabilities directly. Lighter colors correspond to lower values of the spline contribution, while darker and warmer colors indicate higher values, that is, regions of the exposure space where the joint effect of intensity and duration increases the model-predicted risk on the link scale. Differences in color therefore reflect changes in the shape and strength of the estimated interaction surface across alternative preprocessing and modelling specifications. The gray dots represent the observed data and are included to indicate whether data support is available in that area. Each graph overlays contour surfaces derived from alternative preprocessing and modelling choices.

For the oral cavity cancer site, the contour reveal several consistent patterns. At

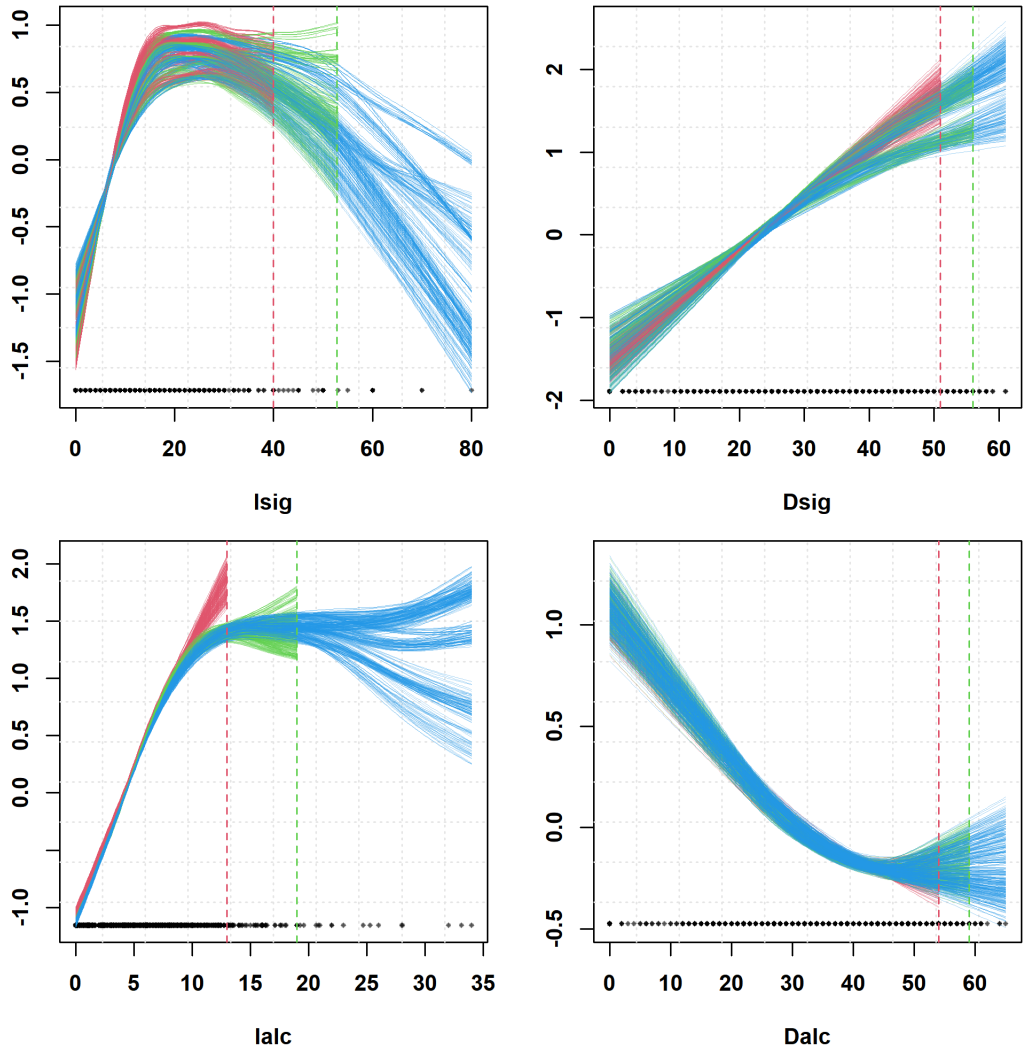


FIGURE 2.8: Stability of univariate spline based exposure–response curves across preprocessing and analytical choices for the esophageal cancer site. Red lines correspond to models using a 95th-percentile cutoff on the exposure variable, green to a 99th-percentile cutoff, and blue to no cutoff.

low levels of smoking intensity (fewer than 10 cigarettes per day), smoking duration contributes little to overall risk, which remains relatively stable. As smoking intensity increases, however, the combined effect of intensity and duration becomes more pronounced. In the upper quantiles of the exposure distribution, some specifications show an apparent decline in risk, a pattern that is not epidemiologically plausible and is likely driven by sparse data at extreme exposure levels. This suggests that, for bivariate spline modelling of smoking consumption, trimming the exposure distribution at the 95th percentile yields more stable and interpretable estimates. This issue is less pronounced for alcohol exposure, where both intensity and duration values are more compactly distributed. In particular, alcohol duration has limited influence and this is clearly visible in the

estimated surfaces that are frequently almost parallel to the duration axis, indicating that duration does not give an important contribution once intensity is accounted for.

For the laryngeal cancer site, smoking intensity and duration jointly contribute to substantial increases in predicted risk. Unlike the other cancer sites, smoking duration exhibits a clear effect even at lower intensity levels, indicating that it plays a more prominent role in laryngeal carcinogenesis. Regarding alcohol exposure alcohol intensity has the strongest influence on predicted risk, while alcohol duration contributes very little, as indicated by contour lines running nearly parallel to the duration axis. The apparent decline in risk beyond approximately 15 units of alcohol intensity is driven by sparse data in that region of the exposure distribution, as evidenced by the absence of gray data points.

For the esophageal cancer, smoking intensity plays a dominant role in shaping risk; even at low to moderate exposure levels, increases in intensity lead to significant increases in predicted risk, while smoking duration has comparatively little influence. This is reflected in the contour lines, which are almost vertical. For alcohol exposure, intensity again emerges as the primary driver of risk, with duration contributing only weakly. The contours for alcohol intensity show strong curvature and steep gradients even at relatively moderate levels, consistent with a nonlinear dose–response relationship. Unlike smoking intensity, however, the distribution of alcohol exposure is more compact, leading to less instability at the upper end of the range. Overall, the esophageal cancer results reinforce the central role of exposure intensity (both for smoking and alcohol) in determining risk, while duration appears to play only a secondary role.

To further validate the necessity of spline based modelling, we examined diagnostic plots of linear predictor estimates across selected exposure ranges for each model specification shown in Figure 2.12, Figure 2.13 and Figure 2.14. For each exposure variable, linear predictor values were computed on a grid of evenly spaced points ranging from 5 to 35 units in 5-unit increments and 5 to 15 units in 7-unit increments for alcohol intensity (due to its narrower exposure range). Under a truly linear relationship, these estimates appear as equidistant points while, deviations from equal spacing reveals nonlinear patterns.

Across all three cancer sites, alcohol intensity exhibits pronounced nonlinearity at higher exposure values; the spacing of linear predictor estimates becomes increasingly irregular at higher exposure levels, indicating that the dose-response relationship is not linear and accelerates at elevated consumption levels. This pattern directly justifies the necessity of spline-based flexible modelling for this exposure variable. For oral cavity and laryngeal cancer, smoking intensity demonstrates strong nonlinearity across the

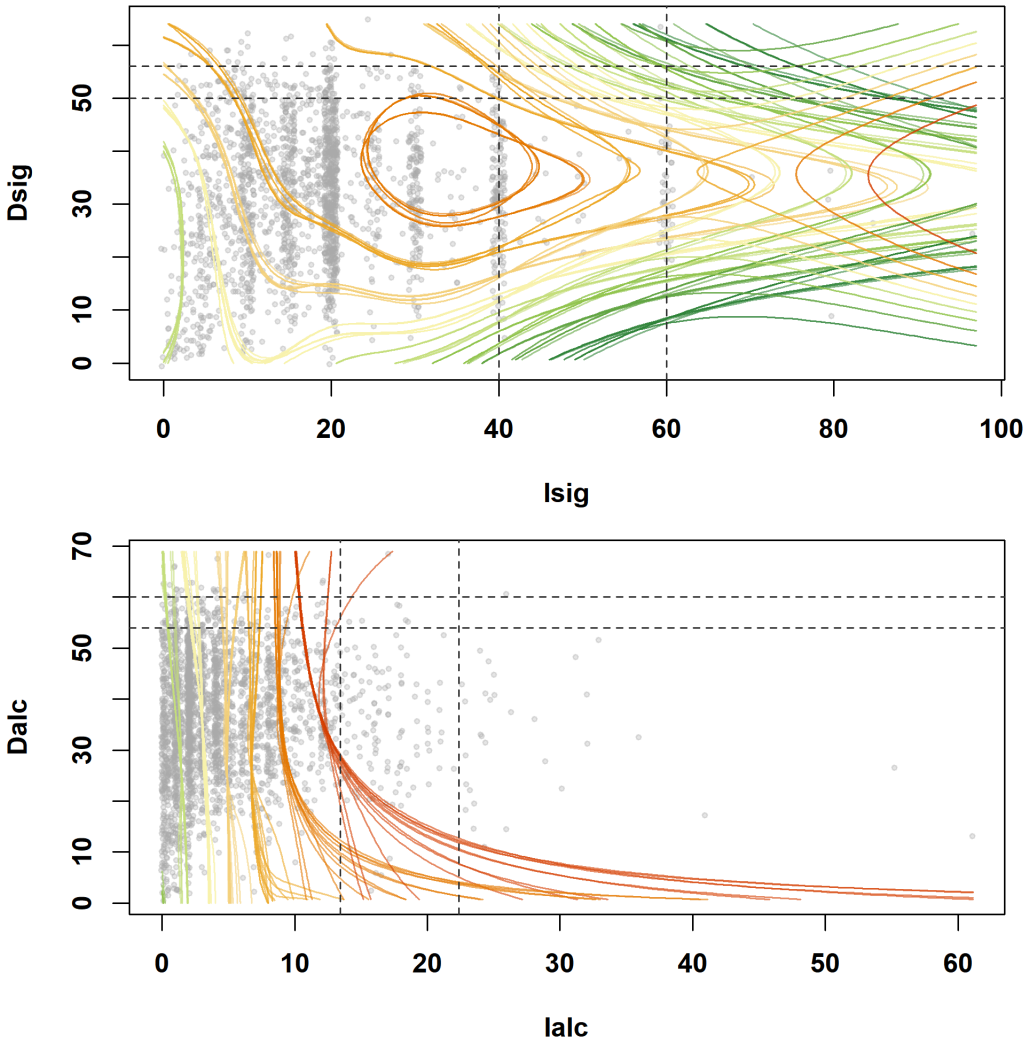


FIGURE 2.9: Bivariate spline contour estimates of predicted probabilities for the joint effects of smoking (top) and alcohol exposure (bottom) for oral cavity cancer. Each contour set corresponds to a different modelling or preprocessing specification while the gray dots are representing the observed data across the exposure space.

exposure gradient. Additionally, when bivariate splines are applied for smoking, the estimated linear predictor values appear more compact across all exposure levels. In contrast, both smoking and alcohol duration generally exhibit much more linear patterns across specifications. The predictor values remain relatively evenly spaced across the examined ranges, supporting the adequacy of linear modelling for these variables. A notable exception occurs in some esophageal cancer specifications, where alcohol duration shows slight nonlinearity, indicating that flexible modelling may offer improvements for this cancer site as well. Overall, these diagnostic plots provide complementary evidence to what previously highlighted.

Finally, in the end we proceed with the validation of the multiverse. To do so, we

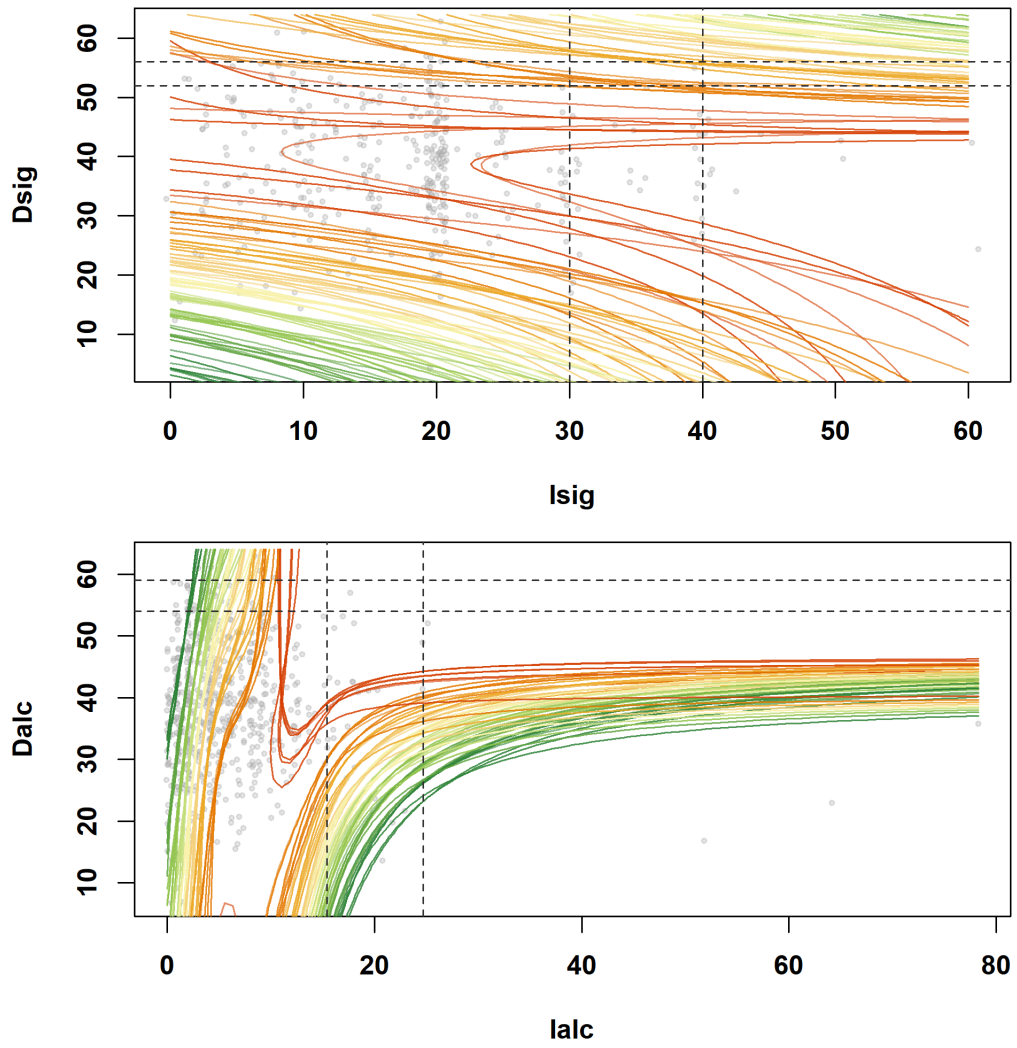


FIGURE 2.10: Bivariate spline contour estimates of predicted probabilities for the joint effects of smoking (top) and alcohol exposure (bottom) for laryngeal cancer. Each contour set corresponds to a different modelling or preprocessing specification while the gray dots are representing the observed data across the exposure space.

conducted a 10 fold cross-validation analysis for all cancer sites for all considered analytical choices. For each combination of spline choices and preprocessing decisions, we computed three classification metrics that are accuracy, sensitivity, and specificity, summarized in the heatmaps shown in Figures 2.15, 2.16 and 2.17. In each heatmap, color intensity represents the magnitude of the corresponding performance metric, with lighter shades indicating lower values and darker shades indicating higher values. The color scale ranges from 0 to 1, consistently across all panels, so that darker cells correspond to better predictive performance in terms of the metric under consideration. This common scale allows direct comparison of performance across modelling specifications within and between panels.

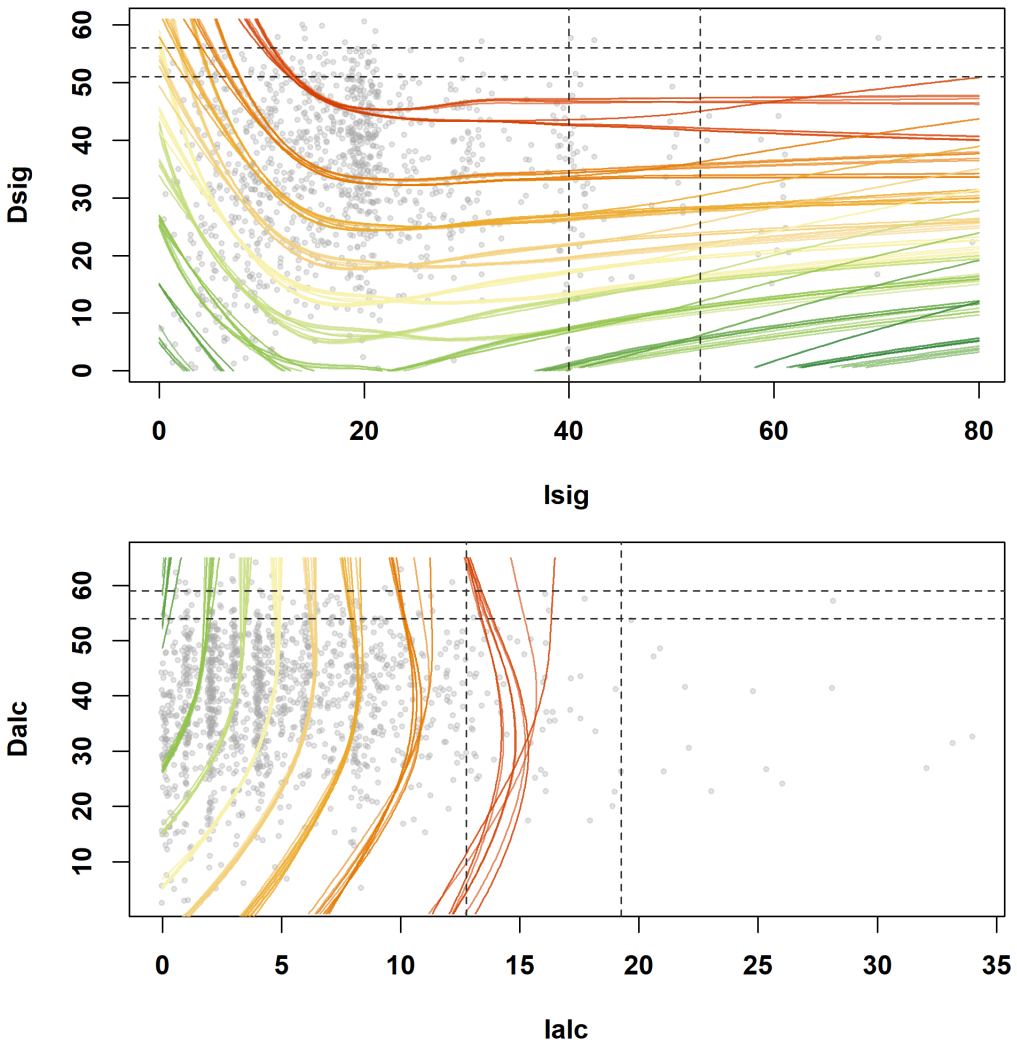


FIGURE 2.11: Bivariate spline contour estimates of predicted probabilities for the joint effects of smoking (top) and alcohol exposure (bottom) for esophageal cancer. Each contour set corresponds to a different modelling or preprocessing specification while the gray dots are representing the observed data across the exposure space.

Across all specifications, for oral cavity cancer, accuracy remains quite stable. Despite substantial variation in functional form as well as differences in preprocessing choices, accuracy remain consistent. This indicates that, for oral cavity cancer, the overall predictive structure is not strongly influenced by modelling decisions, an important result given the large number of alternative specifications explored. A more differentiated pattern emerges when examining sensitivity. The most visible pattern is the sharp reduction in sensitivity observed when alcohol intensity is truncated at the 95th percentile. This cut removes a substantial portion of higher-intensity drinkers, leading the models to perform markedly worse at detecting true cases. Conversely, models that incorporate spline terms for smoking and alcohol intensity generally achieve slightly higher sensitivity,

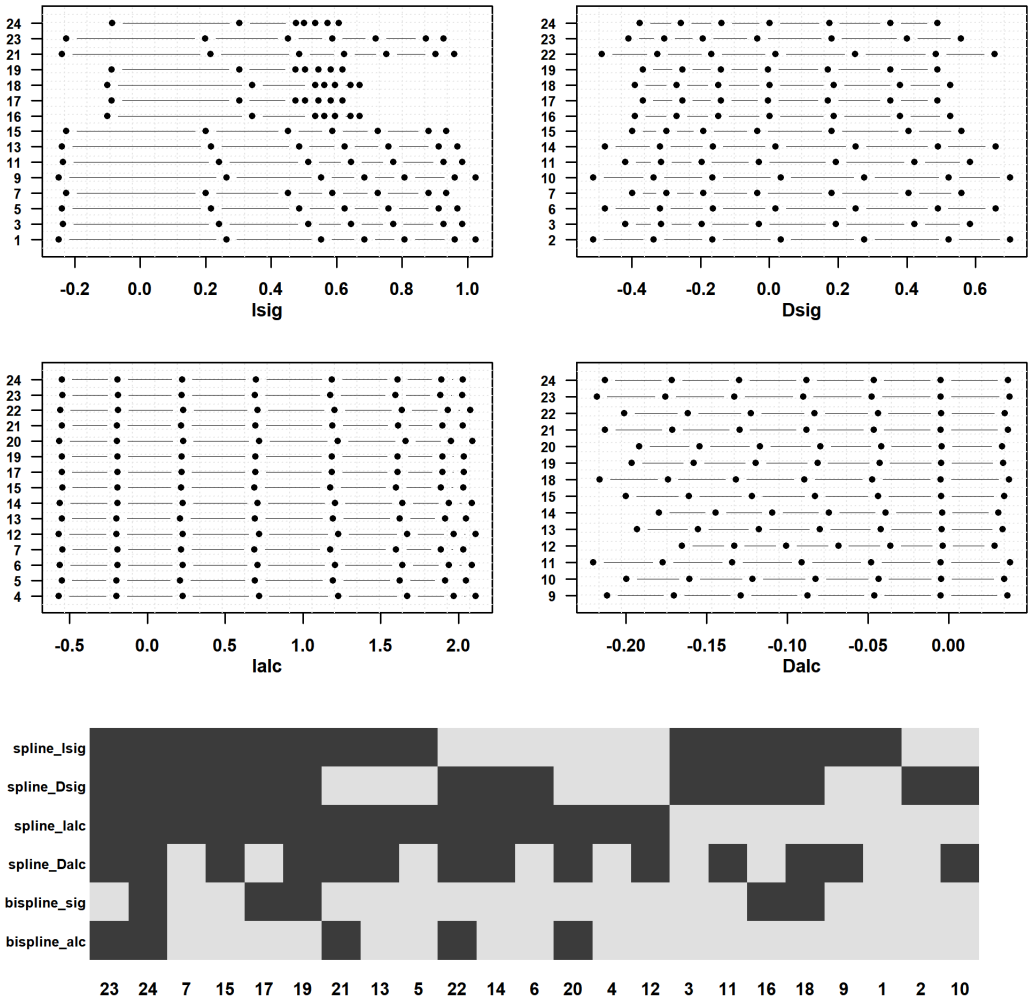


FIGURE 2.12: Assessment of linearity in exposure–response patterns (upper panels) and corresponding model-specification heatmap indicating spline use for each exposure variable (bottom panel), for the oral cavity cancer site.

reflecting their ability to capture the non-linear dose–response patterns identified earlier. In contrast, specificity remains nearly invariant across the multiverse demonstrating a good ability ability to correctly classify non-cases across the entire multiverse. A small increase in specificity appears when alcohol intensity is cut at the 95th percentile, but this improvement comes at the cost of the strong decline in sensitivity noted above.

Cross-validation for laryngeal cancer shows that accuracy is also quite stable across specifications, although is more variable than for the oral cavity. Sensitivity displays stronger differences, with clearly better performance when spline terms for alcohol intensity are included, reflecting the pronounced nonlinearity of the alcohol effect. Specificity, by contrast, remains consistently high across all models, indicating that modelling choices mainly affect case detection rather than false positive detection.

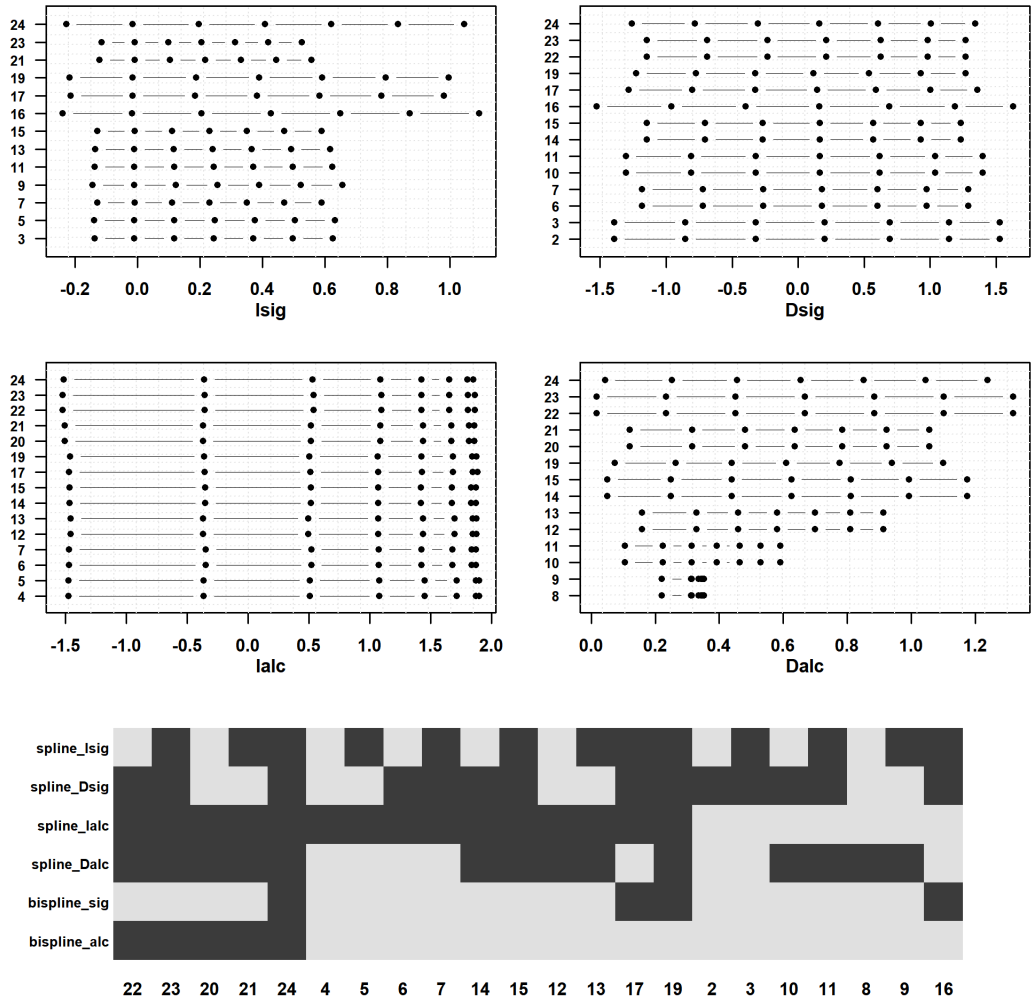


FIGURE 2.13: Assessment of linearity in exposure–response patterns (upper panels) and corresponding model-specification heatmap indicating spline use for each exposure variable (bottom panel), for the laryngeal cancer site.

For esophageal cancer, cross-validation results show that predictive performance is more sensitive to modelling choices than for the other cancer sites but still quite high. Higher values are obtained when spline terms are applied to smoking intensity. Sensitivity is lower as for sites and, using spline terms on alcohol intensity while avoiding cuts on 95th percentile for this exposure leads to better results for the metrics. Specificity remains more uniform showing minor fluctuations across specifications but remaining overall high.

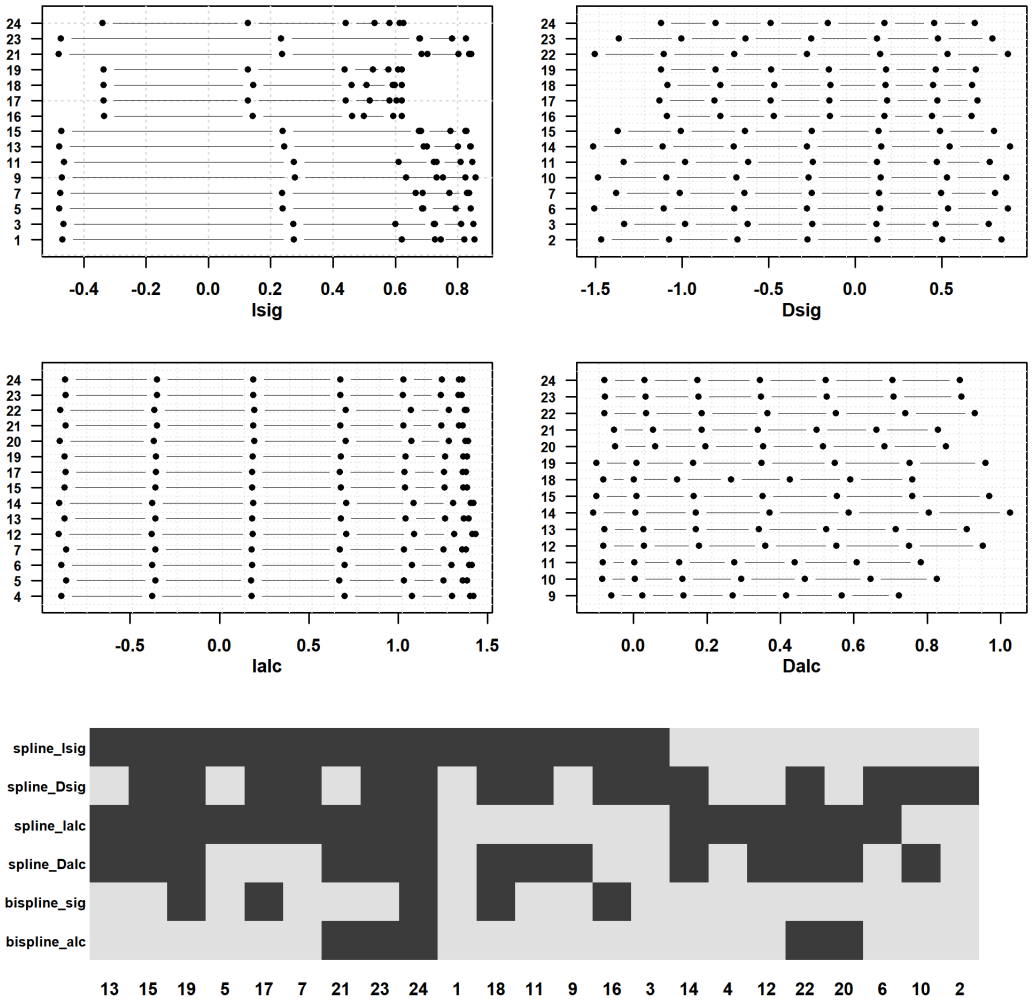


FIGURE 2.14: Assessment of linearity in exposure–response patterns (upper panels) and corresponding model-specification heatmap indicating spline use for each exposure variable (bottom panel), for the esophageal cancer site.

2.5 Conclusions

In this research work, we aimed at addressing the impact of analytical choices on epidemiological risk estimation through a graphical framework based on the ideas behind MA. We explicitly considered a wide range of modelling scenarios that are both statistically grounded and substantively reasonable, and systematically assessed the robustness of empirical findings across these alternative specifications. In particular, we focused on different choices related to data preprocessing and to the modelling of non-linear exposure–response relationships.

Overall, the conclusions regarding risk estimates, expressed in terms of ORs, remained qualitatively robust across the multiverse. However, the analysis revealed a clear

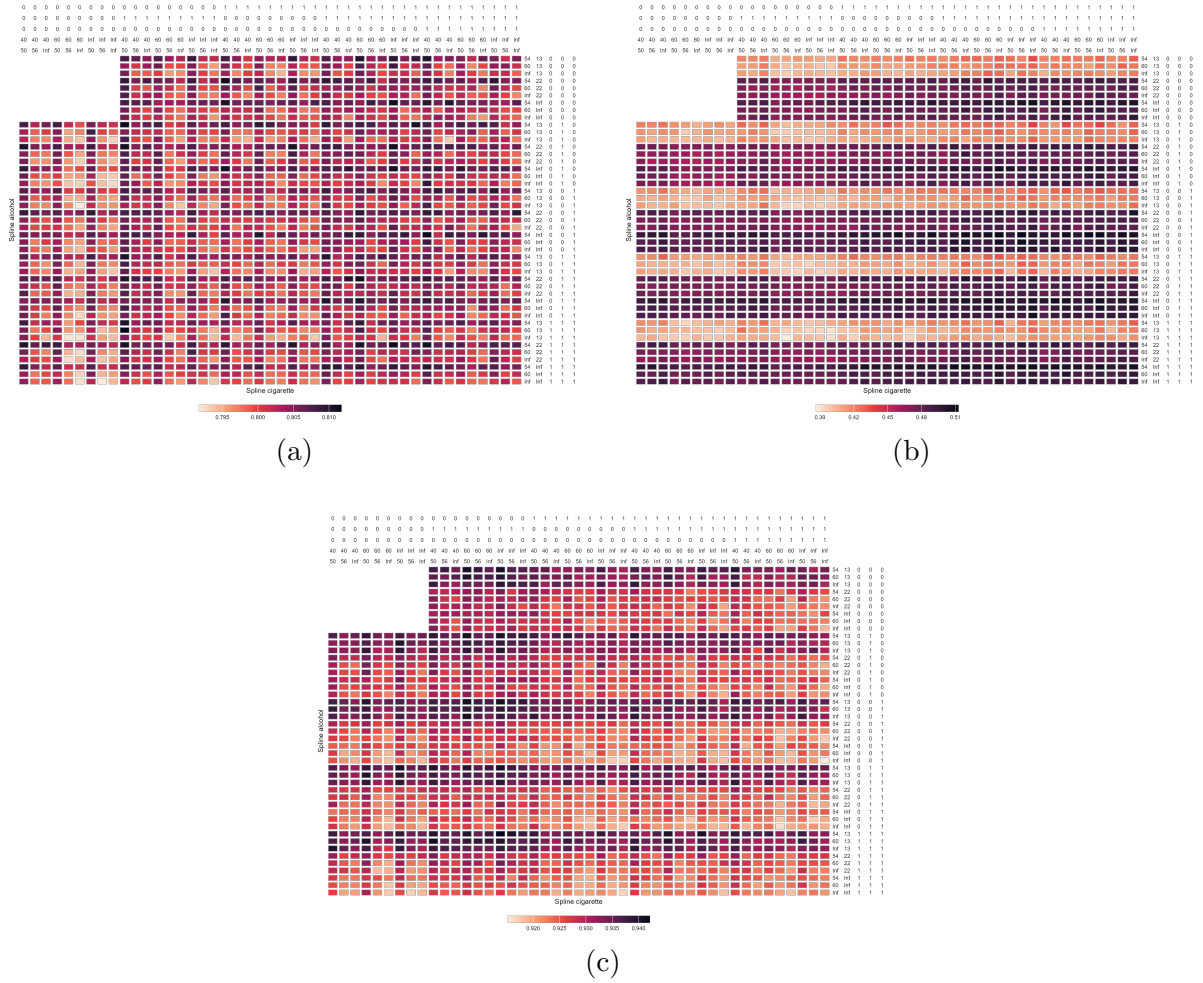


FIGURE 2.15: Cross-validated predictive performance in terms of accuracy (a), sensitivity (b), and specificity (c) for oral cavity cancer across the entire modelling multiverse.

quantitative dependence of the estimates on the chosen specifications. In the empirical application, analyzing the joint effects of smoking and alcohol consumption, in terms of both duration and intensity, on the risk of UADT cancers, pre-processing decisions appeared to have an important impact on risk estimate sensitivity. Specifically, we showed that excluding percentile cut values when they coincide with digit-preference values can lead to a substantial increase in estimated risk.

Furthermore, the assessment of spline based models used to capture nonlinear exposure–response relationships highlighted the use of spline terms appears to be less useful under certain preprocessing choices. This seems to be driven by the removal of extreme exposure values, where spline models are typically most informative. These findings highlight not only the sensitivity of risk estimates to individual analytical decisions, but also the presence of meaningful interactions between different classes of analytical choices. The use of spline terms proved particularly useful when modelling the intensity of alcohol consumption across all cancer sites, evidencing the non linear nature of this exposure

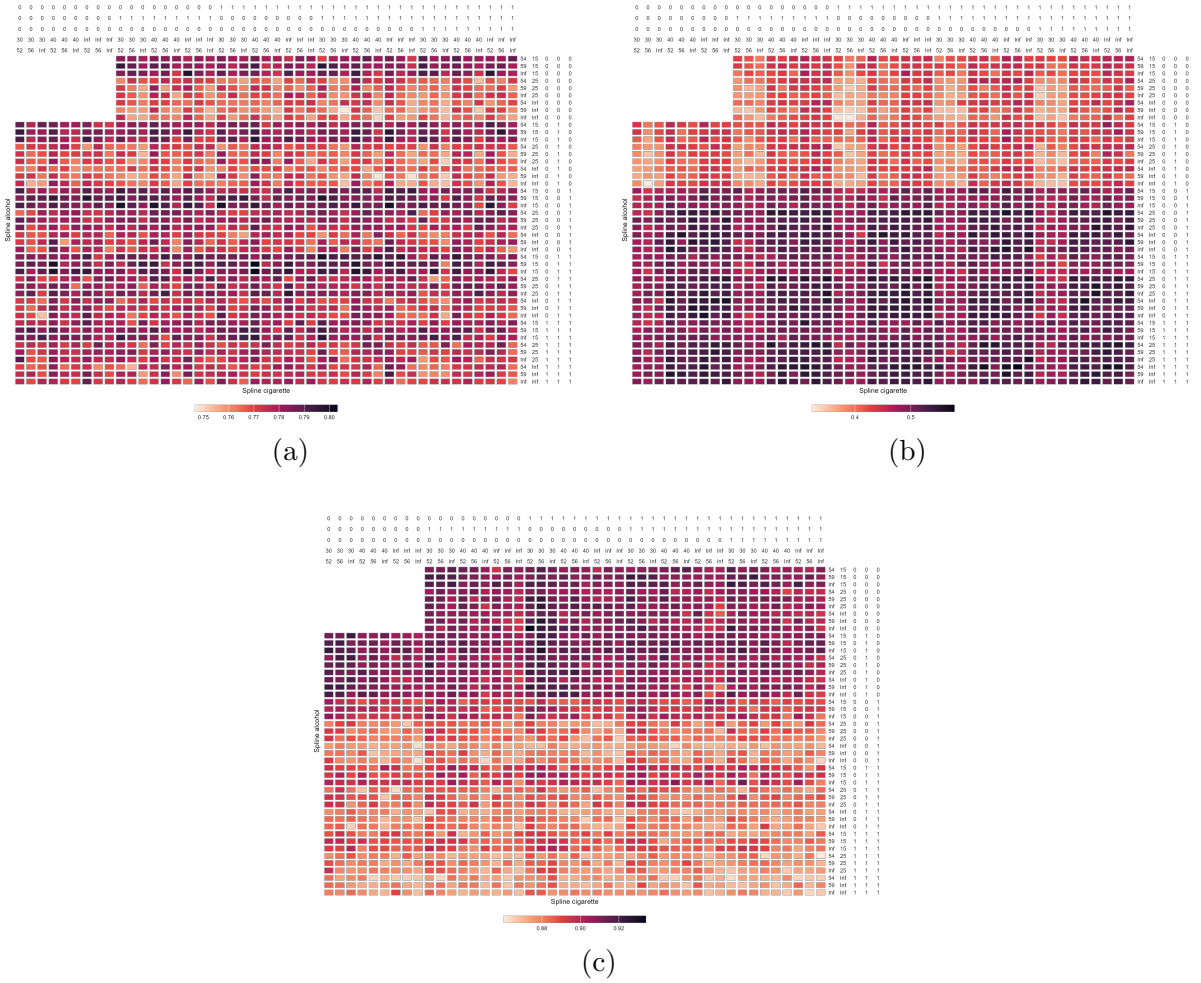


FIGURE 2.16: Cross-validated predictive performance in terms of accuracy (a), sensitivity (b), and specificity (c) for laryngeal cancer across the entire modelling multiverse.

and providing relevant epidemiological insights. For smoking exposure, the analyses indicated a more linear exposure–response pattern compared to alcohol consumption in laryngeal cancer, while for other cancer sites the relationship appeared to be non-linear. Nevertheless, when evaluating the benefits relative to the baseline model with all linear effects, the inclusion of spline terms for smoking intensity did not appear to be relevant in improving model fit in terms of ΔAIC.

Regarding duration of both alcohol and smoking, the exposure–response relationships appeared to be largely linear across most cancer sites. Regarding bivariate splines specifications were generally less influential in improving model fit; however, they offered valuable interpretative information regarding the limited reversibility of alcohol-related risk. Specifically, the results suggest that the effect of alcohol consumption on cancer risk is driven primarily by intensity rather than duration, such that similar levels of risk are observed for individuals with comparable consumption intensity regardless of whether exposure occurred over shorter or longer time periods.

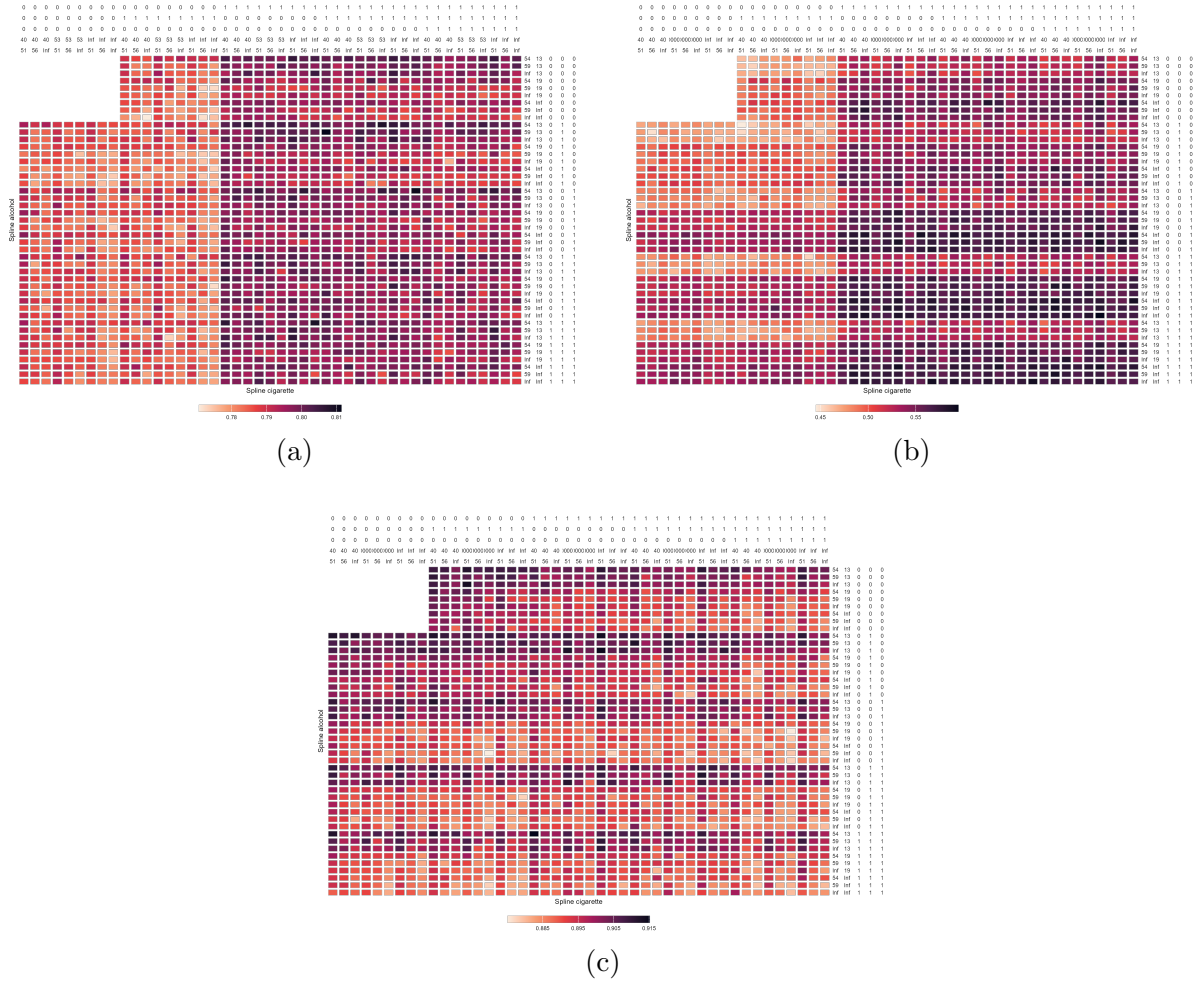


FIGURE 2.17: Cross-validated predictive performance in terms of accuracy (a), sensitivity (b), and specificity (c) for esophageal cancer across the entire modelling multiverse.

Overall, the findings highlight the importance of carefully characterizing exposure–response relationships in epidemiological studies by jointly considering preprocessing and analytical decisions. This work emphasizes the need to make reasonable analytical choices without rigidly adhering to a single prespecified analytical path, instead considering multiple plausible specifications, as the choices made during data analysis may meaningfully influence overall conclusions and have direct implications for the interpretation of risk estimates and public health recommendations. From an epidemiological perspective, this work confirms the importance of adequately accounting for alcohol consumption and smoking exposure in order to prevent the development of UADT cancers, thereby contributing to and reinforcing the existing literature of epidemiological evidence.

Chapter 3

Discovering heterogeneous causal effect of AI and ML usage on sustainable practices across European firms

3.1 Introduction

The last decade has been highlighted as the era of the fourth industrial revolution (Industry 4.0), taking us into a world rapidly moving toward a more digital and data-centric environment. This transition is of particular interest to enterprises, where the need to adopt new digital technologies seems fundamental for business growth and competitiveness. As part of the Industry 4.0 revolution the adoption of technologies like Artificial Intelligence (AI) and machine learning (ML) have emerged as critical components of firms digital transformation journey. Their diffusion across industries is redefining various aspects like production, decision making and innovation processes (McElheran *et al.*, 2024).

Despite the strategic and theoretical importance of these technologies, the real determinants of AI and ML adoption across enterprises remains not fully understood. Existing research points to several key determinants of adoption. Financial resources and digital readiness emerged as particularly influential factors (Kelly *et al.*, 2023) while, what seems to limit adoption appears to be managers inability to identify and effectively manage potential risks (Canhoto and Clear, 2020). Providing a comprehensive understanding of these adoption determinants remains crucial both for identifying disparities in how

different organizations embrace AI and ML (Iansiti and Lakhani, 2016), and for assessing their broader economic impact on business performance (Brynjolfsson and McAfee, 2014).

Beyond the impact on competitiveness, enterprises are also increasingly expected to leverage the use of digital technologies to contribute to environmental and social sustainability goals. This requires that most enterprises need to reach goals such as carbon neutrality, promote circular production and also foster social inclusion (Barteková and Börkey, 2022). In fact, digital technologies are recognized as possible enablers for achieving these objective since the use of innovative technologies may optimize processes and thus, subsequently, reduce energy use and waste, impacting the environmental aspects.

However, at the same time, the use of digital tools may be in contrast with social sustainability goals. The use of digital tools and data-driven systems raises substantive concerns regarding social and ethical dimensions, particularly around employment displacement, data governance, and algorithmic fairness (Brennen and Kreiss, 2016). The automation of routine tasks through advanced digital tools may displace some occupational categories if the whole process of digitalization is not accompanied by adequate reskilling initiatives. This broader alignment between technological innovation, and environmental and social responsibility emphasize the necessity of leading a responsible digital transformation for businesses, moving beyond pure competitiveness. Achieving this integration requires empirical evidence on how firms can effectively leverage emerging technologies, like AI and ML, simultaneously advance business performance, environmental sustainability, and social well-being. This responsible digital transformation is not just a concern for individual companies but policymakers also need to address it. Finding ways to advance digitalization while protecting social and environmental goals is challenging and requires action at multiple levels. Several major institutions have recognized this challenge and are now prioritizing it. The European Commission, for example, has introduced the so called "twin transition" (European Commission, 2020a), which aims to help enterprises transition toward greater efficiency and environmental sustainability at the same time. This policy shows that digitalization and environmental protection are not competing goals but should be pursued together. This also suggest that digital technologies and sustainability goals reinforce each other. Investing in digital innovation supports environmental performance, while sustainability commitments can accelerate digital adoption. Recent empirical research confirms this relationship, demonstrating that among European SMEs, digitalization and sustainability practices tend to evolve in parallell (Aiello *et al.*, 2025).

Despite these insights, existing research has not yet established causal evidence on

how the usage of AI and ML impacts the adoption of sustainable practices. Thus, in this chapter we move beyond simple associations by leveraging causal inference methods to uncover true causal relationships and identify the heterogeneous treatment effects of AI and ML adoption on the introduction of sustainability-conscious practices among European firms. We focus on four concrete practices that European firms may undertake: (i) recycling or reusing materials, (ii) reducing the consumption of, or impact on, natural resources, (iii) saving energy or switching to sustainable energy sources and, (iv) developing sustainable products or services. To examine these relationships, we employ Bayesian Additive Regression Trees (BART), a machine learning method suited for estimating heterogeneous treatment effects in observational data. Unlike traditional regression models that estimate a single average treatment effect (ATE), BART allows us to capture both aggregate effects and firm-level heterogeneity. We explore differences across firm size, sector, and regional characteristics. Our analysis focuses on European small and medium-sized enterprises (SMEs), which form the backbone of the European economy but face particular constraints in the digitalization process compared to larger firms. SMEs typically operate with limited financial resources, lack digital skills, and may be less organizationally oriented toward change. Given their economic significance, the European Commission has prioritized SME digitalization within its wider twin transition agenda. Understanding how AI and ML usage influences sustainability practices differently across SMEs can therefore provide valuable insights for designing targeted policy interventions to enhance the use of digital technologies and consequently have impact also on firms environmental and social sustainability goals.

3.2 Heterogeneous treatment effect modeling in causal inference

In this work, as in the work presented in Chapter 1, we follow the principles behind the potential outcome framework. Thus, as previously stated, the interest is to study the causal effect of a defined treatment or intervention Z on a defined outcome Y. In what follows the treatment is assumed to be dichotomous $Z \in \{0, 1\}$ and, the outcome Y is now a n-dimensional quantity rather than a time-indexed vector, since no temporal structure is considered in this chapter. For each unit i, the potential outcomes are therefore defined as $\{Y_i(0), Y_i(1)\} \in \mathbb{R}^2$. In this case, as in any observational setting, to guarantee the correct identification of the causal effect it is necessary to satisfy the assumptions of consistency, unconfundedness, overlap and SUTVA (see Section 1.2 for details). Alongside the treatment and outcome variable, we consider also a p-dimensional vector of observed

covariates $X_i \in \mathbb{R}^p$ for each sample unit. Each vector incorporates subject specific characteristics that may be considered as potential confounders, influencing both the likelihood of receiving the treatment and the outcome. In the estimation of heterogeneous causal effects (HTE) the vector of covariates plays a central role since, the individual treatment effect (ITE), is in this case summarized conditional on a specific covariate or subsets of covariates. The most general estimand for the HTE is the conditional average treatment effect (CATE):

$$\tau_c = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] \tag{3.1}$$

and quantifies how the causal effect varies as a function of observed characteristics. Unlike the ATE which is obtained marginalizing over the covariates, CATE allows to examine how the treatment effect varies across different subgroups, defined based on covariates, without imposing parametric restrictions. The CATE provides the most general measure for HTE and from this other relevant quantities can be derived. In CATE, we condition upon a specific covariate value while, if we want to condition upon a more aggregated level we can derive the Group Average Treatment Effect (GATE). In this case, instead of conditioning on a single covariate profile, GATE considers predefined or data-driven groups:

$$\tau_g = \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g] \tag{3.2}$$

where G_i denotes the group membership for unit i. GATEs provide a useful summary of treatment heterogeneity at a more interpretable level. To estimate the CATE recently developed methods rely on non parametric machine learning methods or on meta-learners (Jacob, 2021), which can flexibly capture nonlinearities and interactions in X. These approaches are particularly suited for applications in which treatment heterogeneity is expected to play a substantive role due to strong structural differences across population subgroups. Recent years have also witnessed a rapid expansion of methods for HTE leveraging the representational power of large language models (LLMs) and other deep learning structures. Recent work from Athey *et al.*, 2024 shows that using large language model on tabular data converted into text achieves superior predictive performance for complex discrete outcomes in high-dimensional covariate spaces, illustrating the growing potential of this models in causal and predictive settings. Against this, in this work we will rely on using Bayesian Additive Regression Tree (BART) model that is a well-balanced choice for the present application. Our setting involves structured, moderate-dimensional survey data from European SMEs, where the primary goal is

to recover interpretable heterogeneity in treatment effects across firm size, sector, and region — rather than to estimate a single average parameter or to handle unstructured high-dimensional inputs. The models Bayesian regularization and natural posterior uncertainty quantification achieve a practical balance: sufficient flexibility to capture nonlinearities and interactions in the data, without being computationally too complex.

3.2.1 Bayesian Additive Regression Tree

Bayesian Additive Regression Trees (BART), were introduced by Chipman *et al.* (2010) and subsequently extended for causal inference by Hill (2011) and Hahn *et al.* (2020) among others. BART is particularly well suited for estimating heterogeneous treatment effects in observational data because it captures nonlinearities and interactions without requiring an explicit model specification. More formally, it is a nonparametric ensemble method which makes no assumption about the functional form of the relationship between covariates and potential outcomes (Hill *et al.*, 2020). Formally, in BART we assume that the outcome is directly related to an unknown function of covariate set

$$Y_i = f(X_i) + \varepsilon_i$$

where, the unknown function of X_i can be expressed as a sum of many small regression trees

$$f(X_i) = \sum_{j=1}^{J} g_j(X_i)$$

where all regression tree function $g(\cdot)$ explains a small part of the whole heterogeneity providing then a flexible tool that can easily capture complex interactions and nonlinearities. A key point in BART is that it employs priors to constrain the size and depth of trees (so they remain small), preventing overfitting. Posterior inference is performed via Markov Chain Monte Carlo (MCMC) methods, where the tree is grown, pruned and then updated; specifically, the aim is at improving accuracy but also penalizing complex structures. Using MCMC provides natural uncertainty quantification for predictions which would not be easily achievable for non-Bayesian methods. In causal inference, the function we aim at estimating is the potential outcomes under treatment and control given the set of covariates

$$\mu_1(x) = \mathbb{E}[Y_i|X_i = x, Z_i = 1], \qquad \mu_0(x) = \mathbb{E}[Y_i|X_i = x, Z_i = 0]$$

typically fitting each function separately. From these posterior estimates, BART recovers the ITEs

$$\tau_i = \mu_1(X_i) - \mu_0(X_i)$$

from which then aggregated causal estimands as ATE or CATE are derived. Because BART flexibly adjusts for confounding while allowing treatment effects to vary with covariates, it is especially effective for uncovering heterogeneous treatment effects in observational data, as required in this empirical application.

3.3 Empirical application

In this section, we present an empirical application of causal inference methods to observational data in order to estimate the impact of AI and machine learning adoption on firms' sustainability practices. Because treatment effects may vary across firms, the analysis explicitly considers the presence of heterogeneous causal effects. After preprocessing the data to reduce imbalance between treated and untreated units we employ BART, implemented through the `bartCause` package (Dorie *et al.*, 2020), to estimate both average and conditional treatment effects.

3.3.1 Data

The data used in this study come from the "Flash Eurobarometer 486: SMEs, Start-ups, Scale-ups and Entrepreneurship", a survey conducted in 2020 by the GESIS Institute on behalf of the European Commission (European Commission, 2020b). The original sample consists of 16,365 companies located across 84 European and non-European countries. For the purposes of this analysis, we restrict the sample to micro, small and medium sized enterprises (SMEs) operating in the 27 European Union member states.

Furthermore, firms with missing values on the treatment variable were removed. Remaining missing values were addressed using Multiple Imputation by Chained Equations (MICE). Continuous variables were imputed using Classification and Regression Trees (CART), while categorical variables were imputed using polytomous regression models (see Van Buuren, 2000; White *et al.*, 2011). Multiple imputation was preferred to avoid loss of statistical power and to reduce potential bias arising from non-random missingness. The final dataset includes 8,194 SMEs.

Flash Eurobarometer 486 collects detailed information on digitalization, innovation, sustainability practices, and growth strategies, with a particular focus on European

SMEs. The survey aims to identify the main barriers and challenges firms face when transitioning toward more digital and sustainable business models. Interviews are administered to CEOs or senior decision makers through computer-assisted telephone interviewing (CATI).

In our analysis, the treatment variable is defined as a binary indicator capturing whether the firm uses AI or ML methods. The AI/ML usage variable is derived from Question 23 of the Flash Eurobarometer 486, which asks respondents which digital technologies the firm has adopted, including "Artificial intelligence, e.g. machine learning or technologies identifying objects or persons". The variable was coded as 1 if the firm reported using at least one of these technologies and 0 otherwise. The outcome variables correspond to four sustainability related practices measured in the survey that are respectively (i) recycling or reusing materials, (ii) reducing the consumption of or impact on natural resources, (iii) saving energy or switching to sustainable energy sources, and (iv) developing sustainable products or services. The set of covariates selected as potential confounders includes the firm's year of activity, annual turnover, size, revenue, region, sector of operation, and use of other digital technologies. These covariates were selected because we assume they may simultaneously influence both the firm's use of AI/ML technologies and its propensity to engage in sustainability practices, thus satisfying the criterion for confounding adjustment in observational data. Firm size is defined following the EU Recommendation 2003/361, which classifies enterprises according to staff headcount and annual turnover into micro, small, or medium-sized enterprises. Specifically, we classify firms as: (i) micro enterprises those with staff headcount < 10 and annual turnover ≤ 2 millions of euros, (ii) small firms those with staff headcount < 50 and annual turnover ≤ 1 millions of euros and finally (iii) medium enterprises those with staff headcount < 250 and annual turnover ≤ 50 millions of euros. Firms falling outside these thresholds were excluded from the analysis.

Since the study relies on observational data, it is necessary to address potential imbalances between treated and control group prior to estimating causal effects. To do so, we employ propensity score matching (PSM) and all details about the method and results are reported in Appendix C.1.

3.3.2 Results

The empirical analysis provided clear and robust evidence that the adoption of AI and ML technologies has a positive causal impact on the introduction of sustainability practices among European SMEs. The estimated ATEs, reported in red in Figure 3.1, show that AI and ML usage significantly increases firms adoptions of sustainability related

actions across all four domains considered comparing also the obtained results with the causal effect of other digital technologies (ODT) on the same outcome. In general, the strongest impacts are observed for practices aimed at developing sustainable products or services where also the causal effect is slightly higher than for other technologies. In that case the estimated ATE is respectively 0.13 (CI: 0.08 - 0.17) for AI/ML use and 0.12 (CI: 0.08 - 0.17) For the remaining outcomes, the effect is still positive and statistically meaningful since the 95% credible interval is far away from containing the zero.

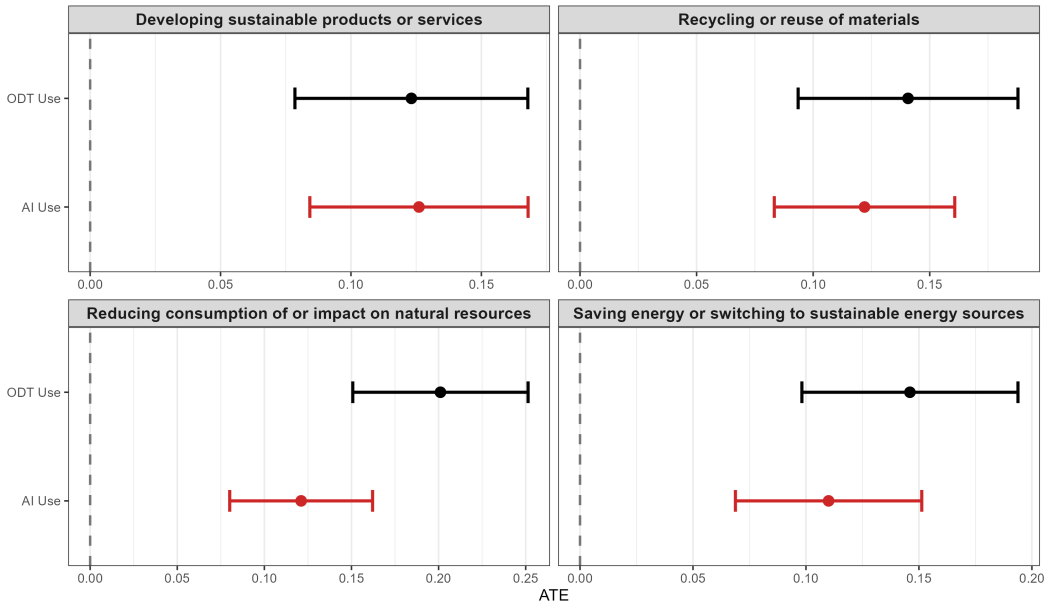


FIGURE 3.1: ATE estimates of AI/ML Use (in red) and Other Digital Technologies (in black) on Sustainability Practices. The figure shows point estimates and 95% credible intervals for ATE.

Regarding the possible presence of heterogeneity on treatment effect we estimated the CATE by firm size reporting results in Figure 3.2. The obtained results indicate that, in general, medium sized enterprises experience lowest improvements in the introduction of sustainability practices following the adoption of AI and ML. Although the differences compared to micro and small firms are small and almost not meaningful, the results suggest that medium sized firms may not be able to turn AI introduction into concrete operational changes as easily as expected, or that they may already have adopted some of these sustainability practices before introducing AI technologies. At the same time, it is important to highlight that the estimated effects remain positive across all size categories. This confirms that AI and ML function are relevant enablers of environmentally oriented actions regardless of firms dimensions. In other words, even firms with more limited resources, such as micro and small enterprises, appear capable of leveraging AI and ML to enhance their sustainability performance.

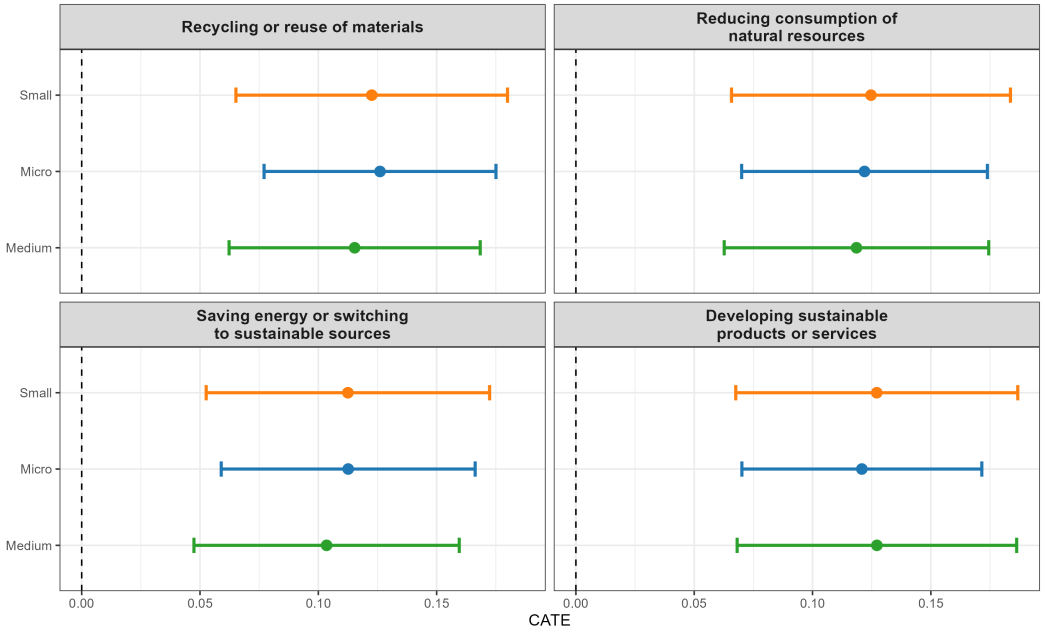


FIGURE 3.2: CATE estimates of AI/ML Use impact on sustainability practices across SME size categories: micro (in blue), small (in orange) and medium (in green). The figure shows point estimates and 95% credible intervals for CATE.

Geographical heterogeneity further characterizes the distribution of the treatment effects. The country level effects displayed in Figure 3.3 show important variation across Europe. In particular, several Southern and Central Eastern European countries exhibit the highest treatment effects across multiple sustainability domains, implying that AI adoption may be particularly relevant in contexts where firms start from lower initial levels of digitalization or sustainability engagement that is at the moment more common in southern Europe. The effect is highly visible in countries as Italy, Greece, Spain and Hungary where digital readiness is still improving and also sustainable developing is not fully embraced yet. In contrast, firms located in Northern and Western European countries tend to exhibit more moderate effects, which may reflect their already well developed digital infrastructures and more advanced sustainability practices. The difference is quite pronounced across all outcomes, except for energy saving practices, where the effect appears slightly more uniform across regions.

To gain more insight about possible heterogeneity and confirm previous findings countries were grouped into three broader macro regions: Central Europe (Germany and Austria), Western Europe (Belgium, the Netherlands, Luxembourg, France and Ireland), Southern Europe (Italy, Spain, Greece, Portugal, Malta and Cyprus), Northern Europe (Denmark, Sweden and Finland), Central Eastern Europe (Poland, Czech Republic, Slovakia and Hungary), Eastern Europe (Bulgaria, Romania, Croatia and Slovenia), and the Baltic States (Estonia, Latvia and Lithuania). The CATE estimates are presented

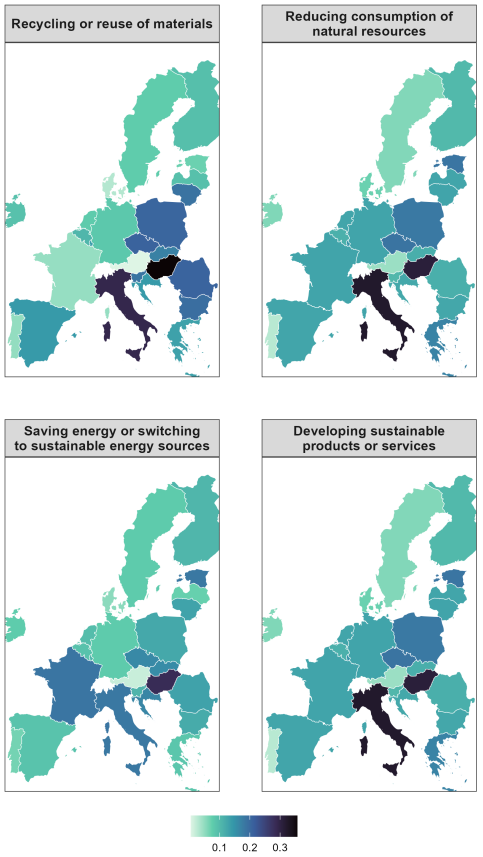


FIGURE 3.3: Country level CATE estimates for AI/ML use effect on sustainability practices. The figure shows the spatial distribution of the CATE across the 27 EU memeber states.

in Figure 3.4. The overall pattern remains consistent: Central Eastern Europe displays the strongest average impacts, followed by Southern and Eastern Europe, while Central, Western, and Northern Europe exhibit smaller but still positive effects. The highest impact is observed for Central Eastern Europe in terms of recycling procedure with a CATE of 0.23.

3.4 Conclusions

This chapter has analyzed the impact of adopting novel digital technologies, as AI and ML, on sustainability practices implemented by European firms, providing empirical evidence on the relationship between digital transformation and environmental performance. The results obtained clearly demonstrate that digital transformation is not just a technical or organizational objective, but rather a strategic driver essential for advancing corporate sustainability.

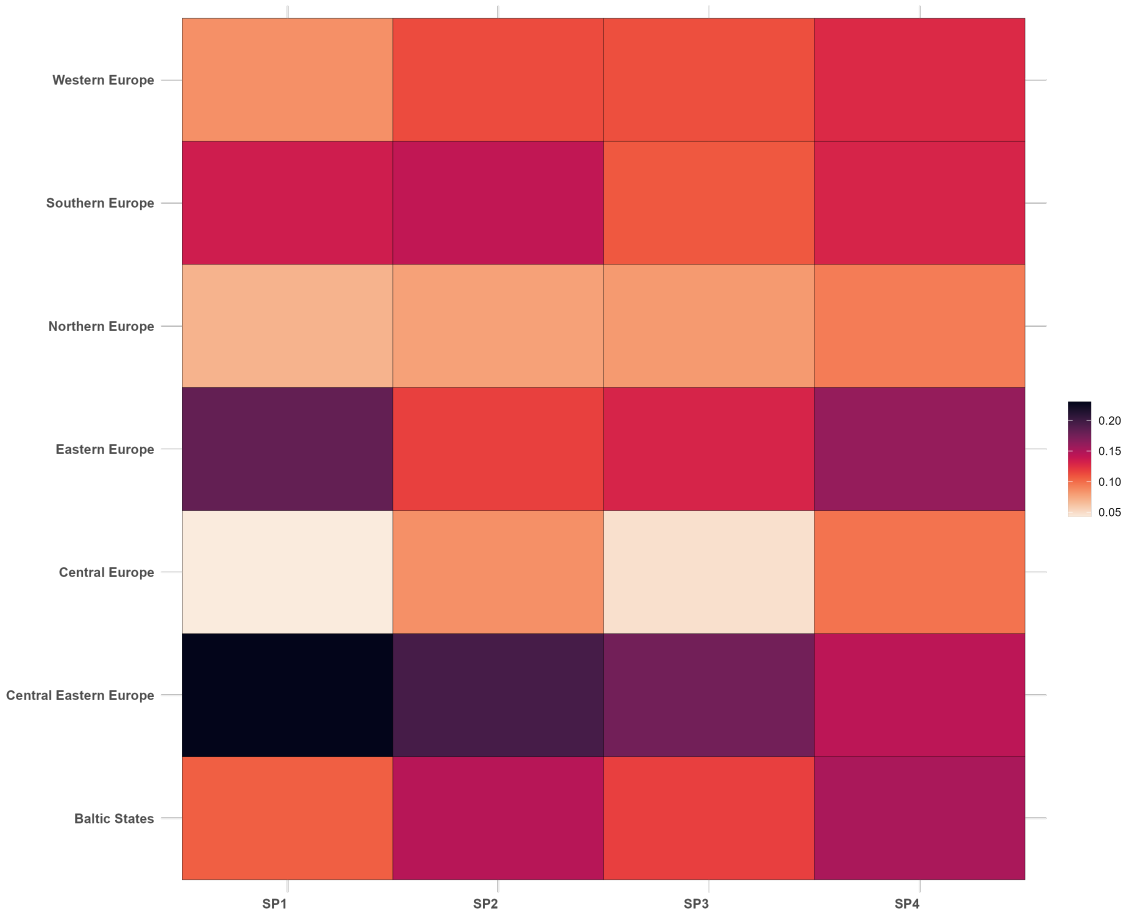


FIGURE 3.4: Macro region level CATE estimates for AI/ML use effect on sustainability practices. The figure shows the distribution of the CATE predefined European regions. Here respectively SP1 refers to recycling or reuse of materials, SF2 to saving energy or switching to sustainable energy sources, SF3 to developing sustainable products or services and, SP4 to reducing consumption of or impact on natural resources.

The empirical analyses conducted have shown that the implementation of AI/ML solutions have positive and statistically significant effects on all sustainability practices considered: from the development of sustainable products, to recycling and material reuse, to reducing consumption and transitioning towards renewable energy sources. This result is particularly relevant in the European context, where the objectives of the European Green Deal and climate neutrality by 2050 require significant accelerations in sustainability efforts.

The positive effects of digitization on sustainability persist regardless of firm size. Although medium-sized enterprises show slightly lower improvements compared to micro and small realities, the effect remains positive across all segments. Furthermore, the analyses have revealed geographical heterogeneity in the effects of AI and ML on sustainability. Central-Eastern European countries (such as Italy, Greece, Spain, and

Hungary) exhibit the largest effects, while Western and Northern European countries show more moderate effects. This reflects the diversity in advancement across EU countries and generates relevant policy implications: European institutions and national governments should prioritise the allocation of public resources, particularly in the form of funding and incentives, towards states in greater need in order to accelerate their digital transition and enhance sustainability.

However, this study is based on cross-sectional data collected at a single point in time, which limits our ability to monitor the temporal dynamics of digitalization effects on sustainability and to distinguish between short-term and long-term impacts. Future research should address these limitations through panel studies that track firms over time. Additionally, implementing statistical clustering methods will allow us to identify specific firm profiles that face greater difficulties in adopting digital technologies for sustainability, as well as to recognise the distinctive organisational, infrastructural, and human capital barriers associated with each cluster. Such analyses will enable the development of more targeted and effective policy interventions based on the heterogeneity of business realities across Europe. Only through such integrated approaches can European economies fully realize the potential of digitalisation as a catalyst for sustainable transition.

Appendix A

A.1 Parallel–trend assumption

Identification of the causal effect relies on the requirement of parallel trend assumption described in Section 1. In the proposed specification, we conditioned upon a low-dimensional latent structure that theoretically we assume are capable of explaining the majority of variability within the data. Formally, for every unit i whose first treatment time is $T_i < \infty$ and for every pre-treatment period $t < T_i$,

$$\mathbb{E}\big[y_{it}(0) \,\big|\, \mathbf{f}_t,\ t < T_i\big] \;=\; \mathbb{E}\big[y_{it}(0) \,\big|\, \mathbf{f}_t,\ T_i = \infty\big]. \tag{A.1}$$

That is, conditional on the latent factors $\mathbf{f}_t$, the untreated outcome of an eventually treated unit is, in expectation, the same as that of a never-treated control. By controlling for potential differences in pre-treatment trends, this condition sub-optimally ensures that any difference observed in block (c) after T_i can be attributed mainly to the intervention.

A.2 Prior distributions for the parameters of interest

In this appendix, the complete set of prior distribution used in the Bayesian dynamic latent factor models introduced in Section 1.3 is specified, excluding the latent factors and factors loadings whose priors where discussed in subsection 1.3.2. All specification are identical for the treated $\mathcal{M}_{Tr}$ and control $\mathcal{M}_C$ sub-models. For each sub-model, let y_{it} denote the observed outcome, $\mathbf{F} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_t^\top) \in \mathbb{R}^{(\mathrm{K}\times\mathrm{T})}$ the latent factors matrix and with $\mathbf{W} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_n^\top) \in \mathbb{R}^{(\mathrm{N}\times\mathrm{K})}$ the corresponding loading matrix. The parameters σ_k^2 and ρ_k denote, respectively, the innovation variance and persistence of the k-th latent factor process, while σ_y^2 represents the observation variance under a Gaussian likelihood, or the overdispersion parameter ϕ under a Negative Binomial likelihood.
For Gaussian outcomes, the observation variance is modeled through the precision

parameter $\upsilon_y = 1/\sigma_y^2$ to which a weakly informative Gamma prior is assigned

$$\upsilon_y \sim \text{Gamma}(a_y, b_y) \tag{A.2}$$

where the rate and the shape parameter are $a_y = 2$ and $b_y = 0.2$, respectively, providing a slightly moderate variances. For count outcomes, where needed, the overdispersion parameter is considered as a fixed value specifically in our case $\phi = 4$. The hyperparameters controlling the dynamic evolution of the factors are given the following priors

$$\rho_k \sim \text{Uniform}(-1, 1), \quad \sigma_k^2 \sim \text{Gamma}(a_k, b_k), \tag{A.3}$$

with rate and shape parameter equal to $a_k = b_k = 0.01$ for each $k = 1, \ldots, K$ factors.

A.3 Auxiliary gradient based sampler

In this appendix the sampling scheme for the Markov chain Monte Carlo used for posterior inference about parameter $\mathbf{F}$ is explained in details.

Algorithm 2 Auxiliary–gradient based update for the latent factors $\mathbf{F}$

Require: Data $\mathbf{Y}$; treatment indicator $\mathbf{D}$; current state $(\mathbf{F}, \boldsymbol{\vartheta})$; step sizes δ

1: Compute log–likelihood at current state: $\ell(\mathbf{Y} \mid \mathbf{F}, \mathbf{D}, \boldsymbol{\vartheta})$
2: Compute likelihood gradient at current state: $\nabla\ell \leftarrow \partial\ell(\mathbf{Y} \mid \mathbf{F}, \mathbf{D}, \boldsymbol{\vartheta})/\partial\mathbf{F}$
3: Draw an auxiliary variable: $\mathbf{z} \sim \mathcal{N}\big(\mathbf{F} + \frac{\delta}{2}\nabla g,\ \frac{\delta}{2}\mathbf{I}\big)$
4: Compute proposal precision: $\mathbf{Q} \leftarrow \big(\mathbf{C}^{-1} + \frac{2}{\delta}\mathbf{I}\big)^{-1}$
5: Compute proposal mean: $\mathbf{m} \leftarrow (\frac{2}{\delta}\,\mathbf{Q}\,\mathbf{z})$
6: Draw a proposal candidate: $\mathbf{F} \sim \text{N}(\mathbf{m}, \mathbf{Q}^{-1})$
7: Compute log–likelihood at proposed state: $\ell(\mathbf{Y} \mid \mathbf{F}^*, \mathbf{D}, \boldsymbol{\vartheta})$
8: Compute likelihood gradient at proposed state: $\nabla\ell \leftarrow \partial\ell(\mathbf{Y} \mid \mathbf{F}^*, \mathbf{D}, \boldsymbol{\vartheta})/\partial\mathbf{F}$
9: Compute Metropolis–Hastings statistic:

$$\log\alpha = \ell(\mathbf{F}^*) - \ell(\mathbf{F}) + \nu(\mathbf{z}, \mathbf{F}^*) - \nu(\mathbf{z}, \mathbf{F}), \qquad \nu(\mathbf{z}, \mathbf{F}) = \big(\mathbf{z} - \mathbf{F} - \tfrac{\delta}{4}\nabla\ell\big)^\top \nabla\ell$$

10: Draw $u \sim \text{Uniform}(0, 1)$; **if** $\log u < \log\alpha$ **then** $\mathbf{F} \leftarrow \mathbf{F}^*$
11: **return** updated $\mathbf{F}$

(Adapt δ to reach ≈50–60 % acceptance.)

A.4 Detailed simulation metric trajectories

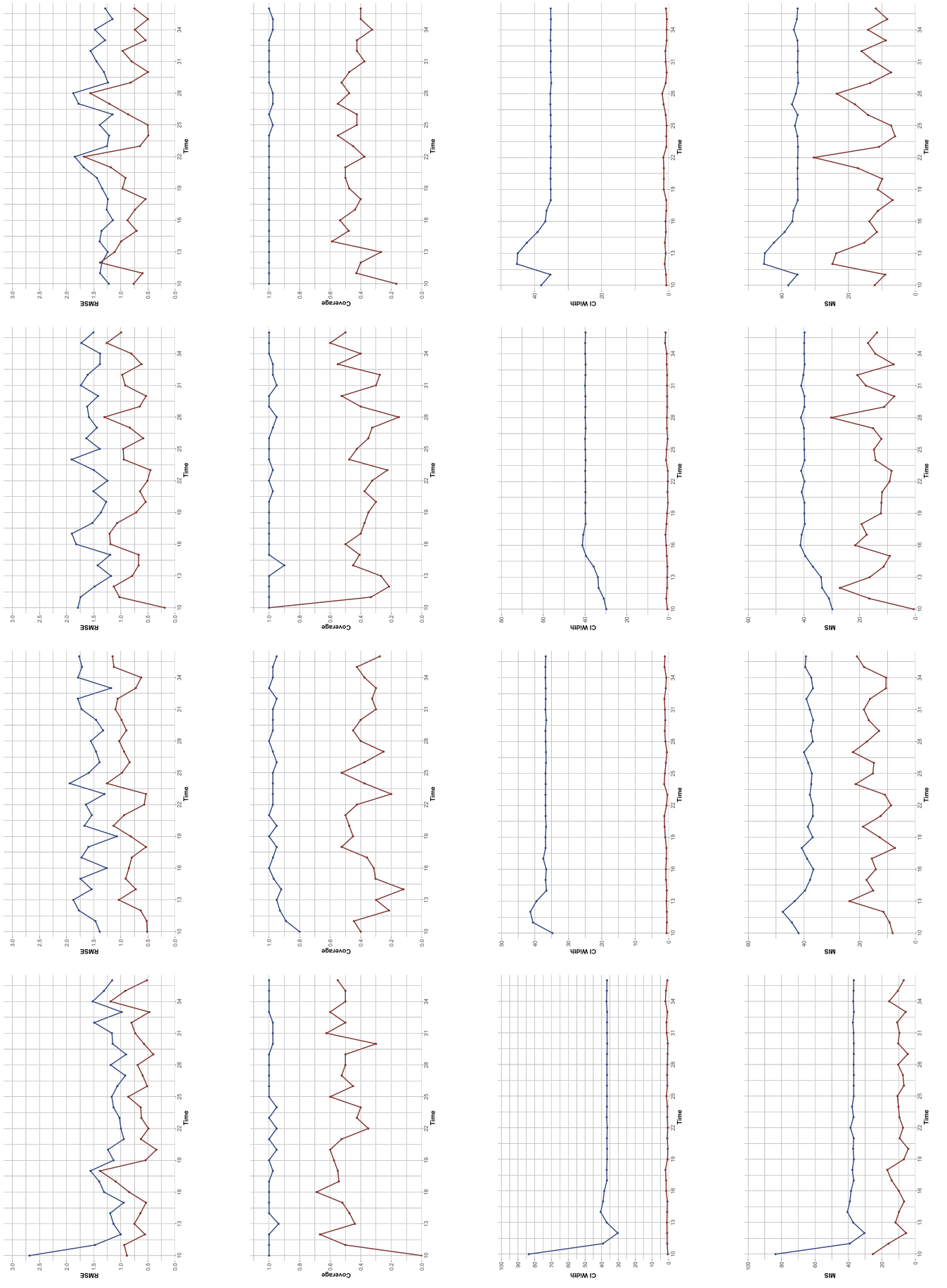


FIGURE A.1: Comparison of ITE estimates under scenarios S1–S4 (columns) for the Gaussian case. Red denotes the proposed mean-based ITE $\tilde{\tau}_{ti}$; blue denotes the estimator $\hat{\tau}_{ti}$. From top to bottom, rows report: RMSE, empirical coverage, average credible-interval width, and Mean Interval Scores (MIS).

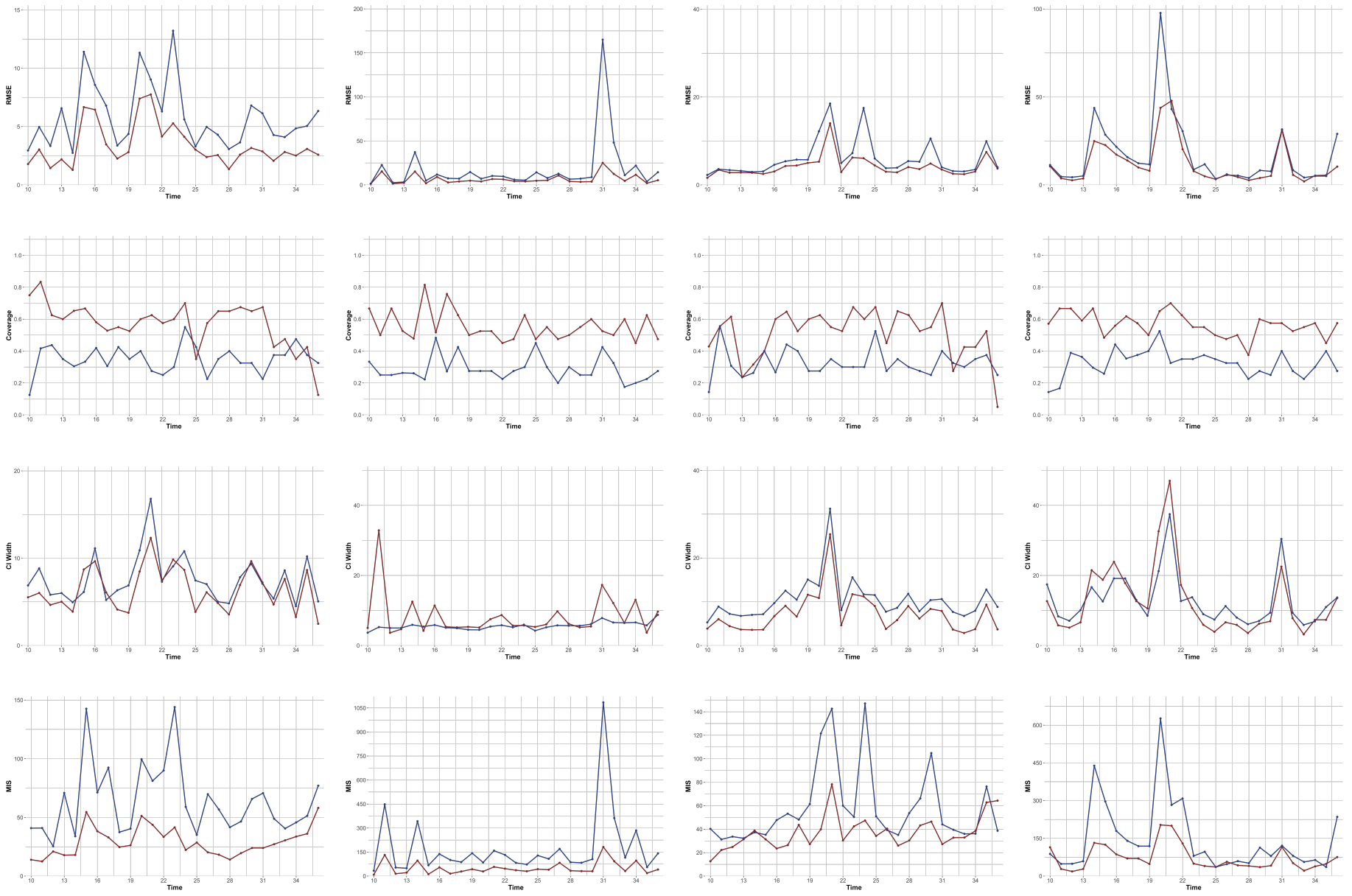


FIGURE A.2: Comparison of ITE estimates under scenarios S1–S4 (columns) for the Negative Binomial case. Red denotes the proposed mean-based ITE $\tilde{\tau}_{ti}$; blue denotes the estimator $\hat{\tau}_{ti}$. From top to bottom, rows report: RMSE, empirical coverage, average credible-interval width, and Mean Interval Scores (MIS).

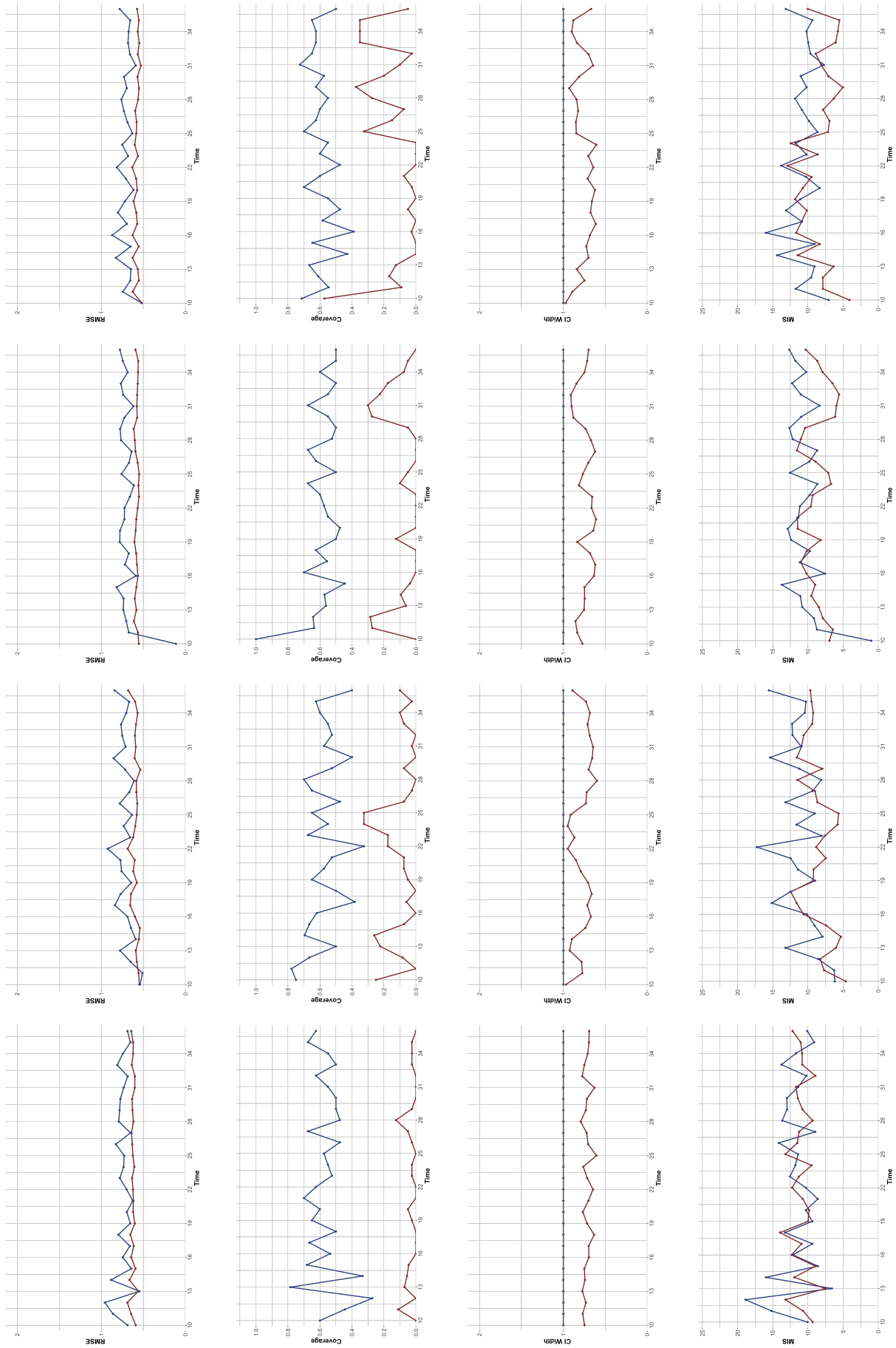


FIGURE A.3: Comparison of ITE estimates under scenarios S1–S4 (columns) for the Bernoullian case. Red denotes the proposed mean-based ITE $\tilde{\tau}_{ti}$; blue denotes the estimator $\hat{\tau}_{ti}$. From top to bottom, rows report: RMSE, empirical coverage, average credible-interval width, and Mean Interval Scores (MIS).

Appendix B

B.1 Preprocessing and Modeling Specifications

In what follows details about the models and analytical choices considered are reported.

TABLE B.1: Preprocessing choices (percentile cut values) applied to each exposure variable by cancer site. For each site and exposure, models were estimated under three alternative trimming options: 95th percentile cut, 99th percentile cut, or no cuts.

Cancer site	**Exposure variable**	**95th percentile**	**99th percentile**	**No cut**
Oral cavity	Smoking intensity (I_{sig})	40	60	∞
Oral cavity	Smoking duration (D_{sig})	50	56	∞
Oral cavity	Alcohol intensity (I_{alc})	13	22	∞
Oral cavity	Alcohol duration (D_{alc})	54	60	∞
Larynx	Smoking intensity (I_{sig})	30	40	∞
Larynx	Smoking duration (D_{sig})	52	56	∞
Larynx	Alcohol intensity (I_{alc})	15	25	∞
Larynx	Alcohol duration (D_{alc})	54	59	∞
Esophagus	Smoking intensity (I_{sig})	40	53	∞
Esophagus	Smoking duration (D_{sig})	51	56	∞
Esophagus	Alcohol intensity (I_{alc})	13	19	∞
Esophagus	Alcohol duration (D_{alc})	54	59	∞

All model specifications are compared to the baseline model, in which all exposure variables enter linearly. The baseline model is:

$$\text{response} \sim s(\text{age}) + \text{sex} + s(\text{education}) + \text{center} + S_{\text{sig}} + S_{\text{alc}} + I_{\text{sig}} + D_{\text{sig}} + I_{\text{alc}} + D_{\text{alc}}.$$

TABLE B.2: List of the 24 alternative model specifications considered in the inferential multiverse. Each specification is expressed in terms of spline terms applied to smoking and alcohol exposure variables.

Model ID	**Specification**
1	$+ s(I_{\text{sig}})$
2	$+ s(D_{\text{sig}})$
3	$+ s(I_{\text{sig}}) + s(D_{\text{sig}})$
4	$+ s(I_{\text{alc}})$
5	$+ s(I_{\text{sig}}) + s(I_{\text{alc}})$
6	$+ s(D_{\text{sig}}) + s(I_{\text{alc}})$
7	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(I_{\text{alc}})$
8	$+ s(D_{\text{alc}})$
9	$+ s(I_{\text{sig}}) + s(D_{\text{alc}})$
10	$+ s(D_{\text{sig}}) + s(D_{\text{alc}})$
11	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(D_{\text{alc}})$
12	$+ s(I_{\text{alc}}) + s(D_{\text{alc}})$
13	$+ s(I_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}})$
14	$+ s(D_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}})$
15	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}})$
16	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + ti(I_{\text{sig}}, D_{\text{sig}})$
17	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(I_{\text{alc}}) + ti(I_{\text{sig}}, D_{\text{sig}})$
18	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(D_{\text{alc}}) + ti(I_{\text{sig}}, D_{\text{sig}})$
19	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}}) + ti(I_{\text{sig}}, D_{\text{sig}})$
20	$+ s(I_{\text{alc}}) + s(D_{\text{alc}}) + ti(I_{\text{alc}}, D_{\text{alc}})$
21	$+ s(I_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}}) + ti(I_{\text{alc}}, D_{\text{alc}})$
22	$+ s(D_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}}) + ti(I_{\text{alc}}, D_{\text{alc}})$
23	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}}) + ti(I_{\text{alc}}, D_{\text{alc}})$
24	$+ s(I_{\text{sig}}) + s(D_{\text{sig}}) + s(I_{\text{alc}}) + s(D_{\text{alc}}) + ti(I_{\text{sig}}, D_{\text{sig}}) + ti(I_{\text{alc}}, D_{\text{alc}})$

In the table $s(\cdot)$ denotes univariate spline terms and $ti(\cdot, \cdot)$ denotes tensor-product bivariate spline terms. To all proposed specifications are also added, as covariates, S_{sig} and S_{alc} (categorical smoking and drinking status: never, former, current), sex, age, education, and center, which are demographic and study-design covariates included in all models as controls.

B.2 Heatmaps for OR for prototype individuals

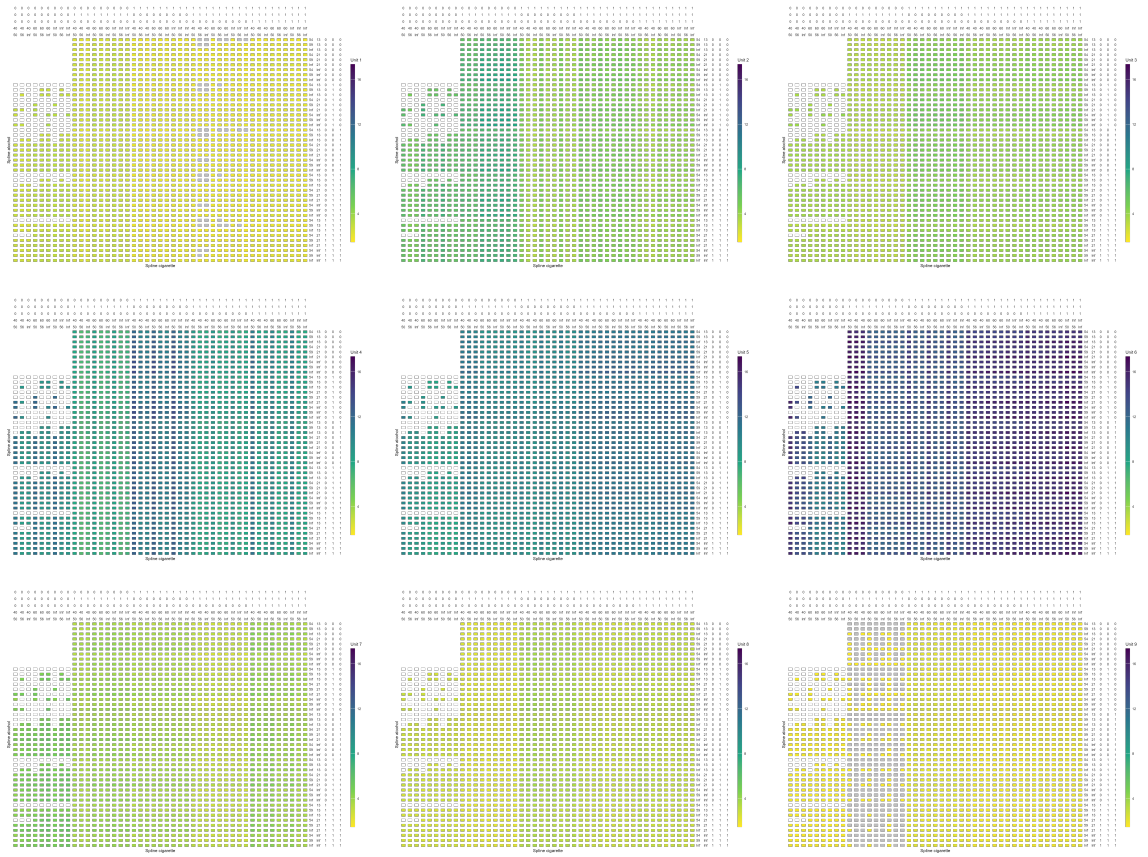


FIGURE B.1: Comparison of OR estimates for all prototype individual across multiverse specifications. The upper line shows the specifications related to smoking habits while the vertical on the right shows the choices regarding drinking habits. For each, first line shows presence of spline for intensity of the exposure, while the second for the duration; third line refers to presence of bivariate splines. Fourth and fifth line indicate the presence of cut-based percentile cuts over the intensity and duration of exposure respectively.

Appendix C

C.1 Propensity score matching

To mitigate the influence of confounding variables when estimating the causal effect of a treatment on an outcome, propensity score matching (PSM) methods are widely employed. The propensity score (PS) is defined as the probability of receiving the treatment conditional on a set of observed baseline covariates. For unit i the propensity score is expressed as

$$e_i(\mathbf{X}) = \mathbb{P}(Z_i = 1|X_i)$$

where Z_i denotes treatment status and X_i represents the vector of observed pretreatment characteristics. In observational studies, the true propensity scores are unknown and must be estimated from the data. This is typically done through regression models using the observed covariates X_i. Once estimated, the propensity score is used to achieve covariate balance, a condition under which the distribution of covariates becomes similar across treated and untreated groups. Achieving covariate balance increases the plausibility that the unconfoundedness assumption holds (for details about PS properties Rosenbaum and Rubin, 1983).

PSM is a widely used non-parametric preprocessing tool that reduces model dependence and improves the comparability between treated and control groups relying on PS. In this study, we employ nearest neighbour matching using the R package `MatchIt` (Ho *et al.*, 2011), to reduce imbalance between firms that use AI/ML technologies and those that do not. The propensity score is first estimated using logistic regression of the treatment indicator on the set of observed pre-treatment covariates. Each treated unit is then paired with the control unit with the closest estimated propensity score, based on a chosen distance metric (for details about matching see Stuart, 2010). It is important to note that PSM provides balance only on the covariates included in the PS model. Thus,

the validity of the method requires that no important confounders remain unobserved, an assumption that may be difficult to guarantee fully. Balance may be assessed by comparing the distribution of covariates across treated and control units before and after matching, often summarized through the standardized mean difference, which captures differences in means but does not guarantee full distributional equivalence. Results are reported in Figure

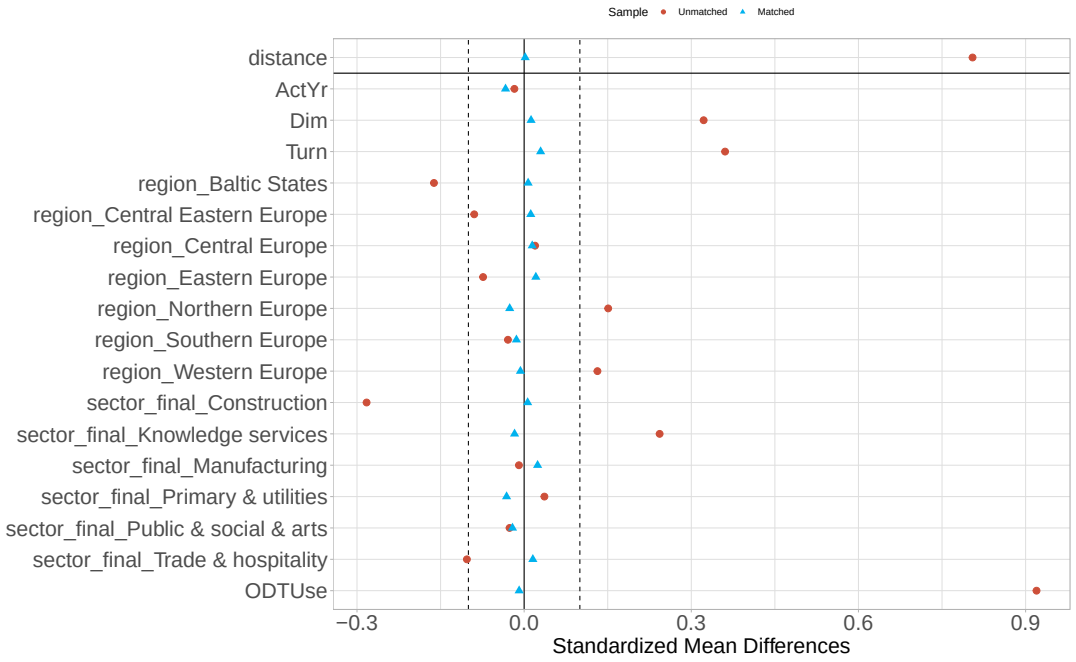


FIGURE C.1: A Comparison in terms of standardized mean difference of covariate balance before and after propensity score matching.

Bibliography

Abadie, A. (2005) Semiparametric difference-in-differences estimators. *The Review of Economic Studies* **72**(1), 1–19.

Abadie, A., Diamond, A. and Hainmueller, J. (2010) Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association* **105**(490), 493–505.

Abadie, A. and Gardeazabal, J. (2003) The economic costs of conflict: A case study of the basque country. *American Economic Review* **93**(1), 113–132.

Advani, A., Elming, W. and Shaw, J. (2023) The dynamic effects of tax audits. *The Review of Economics and Statistics* **105**(3), 545–561.

Aiello, F., Cozzucoli, P. C., Mannarino, L. and Pupo, V. (2025) Bayesian insights on digitalization and environmental sustainability practices. towards the twin transition in the eu. *Business Strategy and the Environment* **34**(1), 417–432.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91**(434), 444–455.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2021) Matrix completion methods for causal panel data models. *Journal of the American Statistical Association* **116**(536), 1716–1730.

Athey, S., Brunborg, H., Du, T., Kanodia, A. and Vafa, K. (2024) Labor-llm: Language-based occupational representations with large language models. *arXiv preprint arXiv:2406.17972* .

Bai, J. and Ng, S. (2006) Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74**(4), 1133–1150.

Baldwin, J. R., Pingault, J.-B., Schoeler, T., Sallis, H. M. and Munafò, M. R. (2022) Protecting against researcher bias in secondary data analysis: challenges and potential solutions. *European Journal of Epidemiology* **37**(1), 1–10.

Barteková, E. and Börkey, P. (2022) Digitalisation for the transition to a resource efficient and circular economy. Technical report, OECD Publishing.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., de Jonge, P., Emerencia, A. C., Epskamp, S. *et al.* (2020) Time to get personal? the impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of psychosomatic research* **137**, 110211.

Ben-Michael, E., Arbour, D., Feller, A., Franks, A. and Raphael, S. (2023) Estimating the effects of a california gun control program with multitask gaussian processes. *The Annals of Applied Statistics* **17**(2), 985–1016.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A. *et al.* (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**(7810), 84–88.

Box, G. E. (1979) Robustness in the strategy of scientific model building. In *Robustness in statistics*, pp. 201–236. Elsevier.

Bravi, F., Lee, Y.-C. A., Hashibe, M., Boffetta, P., Conway, D. I., Ferraroni, M., La Vecchia, C., Edefonti, V., Investigators, I. C., Agudo, A. *et al.* (2021) Lessons learned from the inhance consortium: An overview of recent results on head and neck cancer. *Oral Diseases* **27**(1), 73–93.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* **16**(3), 199–231.

Brennen, J. S. and Kreiss, D. (2016) Digitalization. *The international encyclopedia of communication theory and philosophy* pp. 1–11.

Brynjolfsson, E. and McAfee, A. (2014) *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & company.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T. *et al.* (2016) Evaluating replicability of laboratory experiments in economics. *Science* **351**(6280), 1433–1436.

Canhoto, A. I. and Clear, F. (2020) Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. *Business Horizons* **63**(2), 183–193.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society* **158**(3), 419–444.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010) Bart: Bayesian additive regression trees .

Christiansen, T. G. (2024) Dynamic effects of tax audits and the role of intentions. *Journal of Public Economics* **234**, 105121.

Constantino, S. M., Pianta, S., Rinscheid, A., Frey, R. and Weber, E. U. (2021) The source is the message: the impact of institutional signals on climate change–related norm perceptions and behaviors. *Climatic Change* **166**(3), 35.

Cotter, S. L., Roberts, G. O., Stuart, A. M. and White, D. (2013) Mcmc methods for functions: modifying old algorithms to make them faster. *Statistical Science* .

De Finetti, B. (2017) *Theory of probability: A critical introductory treatment*. John Wiley & Sons.

DeBacker, J., Heim, B. T., Tran, A. and Yuskavage, A. (2018) Once bitten, twice shy? the lasting impact of enforcement on tax compliance. *Journal of Law and Economics* **61**(1), 1–35.

Del Giudice, M. and Gangestad, S. W. (2021) A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science* **4**(1), 2515245920954925.

Di Credico, G., Edefonti, V., Polesel, J., Pauli, F., Torelli, N., Serraino, D., Negri, E., Luce, D., Stucker, I., Matsuo, K. *et al.* (2019) Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer: A bivariate spline model approach. *Oral oncology* **94**, 47–57.

Di Credico, G., Polesel, J., Dal Maso, L., Pauli, F., Torelli, N., Luce, D., Radoï, L., Matsuo, K., Serraino, D., Brennan, P. *et al.* (2020) Alcohol drinking and head and neck cancer risk: the joint effect of intensity and duration. *British journal of cancer* **123**(9), 1456–1463.

Dorie, V., Hill, J. and Dorie, M. V. (2020) Package 'bartcause'. *URL: https://cran. r-project. org/web/packages/bartCause/bartCause. pdf* .

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K. *et al.* (2018) Model averaging in ecology: A review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological monographs* **88**(4), 485–504.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. and Nosek, B. A. (2021) Investigating the replicability of preclinical cancer biology. *elife* **10**, e71601.

European Commission (2020a) New industrial strategy for europe. Technical Report COM(2020) 102 final, European Commission, Brussels. 2020a.

European Commission (2020b) Flash eurobarometer 486: Smes, start-ups, scale-ups and entrepreneurship. Technical Report ZA7637 Data File Version 2.0.0, GESIS Data Archive Cologne; Kantar Belgium, Brussels, Belgium. 2020b.

Fan, J., Xue, L. and Yao, J. (2017) Sufficient forecasting using factor models. *Journal of econometrics* **201**(2), 292–306.

Fisher, R. A. (1925) Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pp. 700–725.

Franceschi, S., Talamini, R., Barra, S., Barón, A. E., Negri, E., Bidoli, E., Serraino, D. and La Vecchia, C. (1990) Smoking and drinking in relation to cancers of the oral cavity, pharynx, larynx, and esophagus in northern italy. *Cancer research* **50**(20), 6502–6507.

Gelman, A. and Loken, E. (2013) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* **348**(1-17), 3.

Geweke, J. and Zhou, G. (1996) Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies* **9**(2), 557–587.

Gobillon, L. and Magnac, T. (2016) Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics* **98**(3), 535–551.

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

Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statistical science* **1**(3), 297–310.

Held, L. (2020) A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society Series A: Statistics in Society* **183**(2), 431–448.

Hill, J., Linero, A. and Murray, J. (2020) Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application* **7**(1), 251–278.

Hill, J. L. (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

Ho, D., Imai, K., King, G. and Stuart, E. A. (2011) Matchit: nonparametric preprocessing for parametric causal inference. *Journal of statistical software* **42**, 1–28.

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U. and Boulesteix, A.-L. (2021) The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. *Royal Society open science* **8**(4), 201925.

Hsiao, C., Ching, S. H. and Ki, S. K. (2012) A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics* **27**(5), 705–740.

Iansiti, M. and Lakhani, K. (2016) The digital business divide. *Analyzing the operating impact of digital transformation. Boston: Harvard Business School Report* .

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Ioannidis, J. P. (2008) Why most discovered true associations are inflated. *Epidemiology* **19**(5), 640–648.

Jacob, D. (2021) Cate meets ml: Conditional average treatment effect and machine learning. *Digital Finance* **3**(2), 99–148.

Kelly, S., Kaye, S.-A. and Oviedo-Trespalacios, O. (2023) What factors contribute to the acceptance of artificial intelligence? a systematic review. *Telematics and informatics* **77**, 101925.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. and Saez, E. (2011) Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica* **79**(3), 651–692.

Leamer, E. E. (1985) Sensitivity analyses would help. *The American Economic Review* **75**(3), 308–313.

Macrì Demartino, R., Egidi, L., Held, L. and Pawel, S. (2024) Mixture priors for replication studies. *arXiv e-prints* pp. arXiv–2406.

Maie, R., Eguchi, M. and Uchihara, T. (2024) Arbitrary choices, arbitrary results: Three cases of multiverse analysis in l2 research. *Research Methods in Applied Linguistics* **3**(2), 100124.

McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L. and Zolas, N. (2024) Ai adoption in america: Who, what, and where. *Journal of Economics & Management Strategy* **33**(2), 375–415.

Micheloud, C., Balabdaoui, F. and Held, L. (2023) Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Statistica Neerlandica* **77**(4), 573–591.

Micheloud, C. and Held, L. (2024) The replication of equivalence studies. *Biometrical Journal* **66**(8), e202300232.

Neyman, J. and Pearson, E. S. (1933) The testing of statistical hypotheses in relation to probabilities a priori. In *Mathematical proceedings of the Cambridge philosophical society*, volume 29, pp. 492–510.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* **349**(6251), aac4716.

Patel, C. J., Burford, B. and Ioannidis, J. P. (2015) Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of clinical epidemiology* **68**(9), 1046–1058.

Pawel, S. and Held, L. (2022) The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(3), 879–911.

Pelucchi, C., Gallus, S., Garavello, W., Bosetti, C. and La Vecchia, C. (2008) Alcohol and tobacco use, and cancer risk for upper aerodigestive tract and liver. *European journal of cancer prevention* **17**(4), 340–344.

Polesel, J., Dal Maso, L., Bagnardi, V., Zucchetto, A., Zambon, A., Levi, F., La Vecchia, C. and Franceschi, S. (2005) Estimating dose-response relationship between ethanol and risk of cancer using regression spline models. *International journal of cancer* **114**(5), 836–841.

Polesel, J., Talamini, R., La Vecchia, C., Levi, F., Barzan, L., Serraino, D., Franceschi, S. and Dal Maso, L. (2008) Tobacco smoking and the risk of upper aero-digestive tract cancers: A reanalysis of case–control studies using spline models. *International journal of cancer* **122**(10), 2398–2402.

Roberts, G. O. and Stramer, O. (2002) Langevin diffusions and metropolis-hastings algorithms. *Methodology and Computing in Applied Probability* **4**, 337–357.

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

Rosenberg, P. S., Katki, H., Swanson, C. A., Brown, L. M., Wacholder, S. and Hoover, R. N. (2003) Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Statistics in medicine* **22**(21), 3369–3381.

Simonsohn, U., Simmons, J. P. and Nelson, L. D. (2020) Specification curve analysis. *Nature human behaviour* **4**(11), 1208–1214.

Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016) Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**(5), 702–712.

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

Stuart, E. A., DuGoff, E., Abrams, M., Salkever, D. and Steinwachs, D. (2013) Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *eGEMs* **1**(3), 1038.

Titsias, M. K. and Papaspiliopoulos, O. (2018) Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(4), 749–767.

Van Buuren, S. (2000) *Multivariate imputation by chained equations: MICE V1. 0 user's manual*. Leiden: TNO.

Vorland, C. J., O'Connor, L. E., Henschel, B., Huo, C., Shikany, J. M., Serrano, C. A., Henschel, R., Dickinson, S. L., Ejima, K., Bidulescu, A. *et al.* (2025) "shaking the ladder" reveals how analytic choices can influence associations in nutrition epidemiology: beef intake and coronary heart disease as a case study. *Critical Reviews in Food Science and Nutrition* pp. 1–16.

White, I. R., Royston, P. and Wood, A. M. (2011) Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine* **30**(4), 377–399.

Wood, S. N. (2001) mgcv: Gams and generalized ridge regression for r. *R news* **1**(2), 20–25.

Xu, Y. (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* **25**(1), 57–76.

Young, C. (2018) Model uncertainty and the crisis in science. *Socius* **4**, 2378023117737206.

La borsa di dottorato è cofinanziata con risorse dell'Unione europea, NextGeneration EU - Piano Nazionale di Ripresa e Resilienza, Missione 4 – Componente 2 – Investimento 3.3 **CUP J92B22001000007**