DOCTORAL THESIS

Statistical methods for longitudinal medical data with applications.

Author:
Martina AMONGERO

Supervisor:
Prof. Mauro GASPARINI

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Pure and Applied Mathematics

Statistics and Data Science Group
Department of Mathematical Science *Giuseppe Luigi Lagrange*

POLITECNICO DI TORINO - UNIVERSITÀ DEGLI STUDI DI TORINO

May 2024

Abstract

In this thesis, we discuss the use of longitudinal data in biostatistics and their analysis, focusing on three specific real cases of study.

Longitudinal data refer to collections of repeated measurements of specific variables of interest at multiple time points. Their analysis offers many advantages: among others, it enables the evaluation of temporal evolutions of quantities of interest (biomarkers, tumor size, daily counts,..), also from an individual perspective, and it provides stronger evidence for causal relationships. Various statistical methods can be used to analyze longitudinal data. They range from generalized mixed-effect models, growth and evolution modeling (often combined with the mixed-effects structures), to time-to-events analyses. However, such statistical methodologies might sometimes involve complicated issues to deal with, especially those related to censoring and missing data problems.

In this work, we present three longitudinal studies. (I) The first one focuses on modeling and forecasting the COVID-19 pandemic in Italy using a newly developed compartmental model called SIPRO. Its analysis shows the necessity of extending the well-known SIR model to account for the asymptomatic part of the population, in order to realistically describe the COVID-19 pandemic. Moreover, it warns about identifiability issues that arise when the extended model is too complicated with respect to the collected information. (II) The second one focuses on longitudinal data from prostate cancer patients and it aims at estimating the optimal time to recommend an expensive examination for prostate cancer patients who presented a resurgence after surgery. In particular, this study highlights that better estimates can be obtained, with respect to logistic models applied so far, using a more complex joint model that incorporates all the patients clinical history. (III) Finally, the third one addresses the practical implementation of pre-existing methodologies discussed in the literature. Specifically, it focuses on adapting one of these methods to account for informative withdrawal in recurrent event problems, with the aim of estimating vaccine efficacy. Based on a real case study, provided by GSK, this work shows how to obtain more reliable estimates in case of missing data due to informative censoring, and warns about numerical issues that can arise during the analyses.

Contents

List of Figures

List of Tables

List of Abbreviations

AECOPD	Acute Exacerbations of Chronic Obstructive Pulmonary Disease
AG	Andersen and Gill model
AIC	Akaike Information Criterion
BCP	Biochemical Persistence
BCR	Biochemical Relapse
CI	Confidence or Credible Interval (depending on the framework)
COPD	Chronic Obstructive Pulmonary Disease
ES	Energy Score
FU	Follow Up
GAM	Generalized Additive Model
GLM	Generalized Linear Model
GS	Gibbs Sampler
IPCW	Inverse Probability Censoring Weigths
IPW	Inverse Probability Weigths
IR	Intensity Ratio
JFM	Joint Frailty Model
KM	Kaplan Meier
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
MLE	Maximum Likelihood Estimator
ODE	Ordinary Differential Equation
PET-PSMA	Positron Emission Tomography with PSMA
PSA	Prostate-Specific Antigen
PSMA	Prostate-Specific Membrane Antigen
PT	Parallel Tempering
SIPRO	Susceptibles Infected Positive Recovered Out
SIR	Susceptible Infected Recovered
VE	Vaccine Efficacy
WAIC	Watanabe–Akaike Information Criterion

List of Symbols

$\hat{\alpha}$	Estimated value for the parameter α
$Ber(p)$	Bernoulli distribution with probability parameter p
$Bin(n, p)$	Binomial distribution with n trials and probability of success p
$\stackrel{D}{=}$	Equality in distribution
$diag(a_1, ..., a_n)$	Diagonal $n \times n$ matrix with elements $(a_1, ...a_n)$ on the diagonal
$Dirichlet(a_1, ..., a_n)$	Dirichlet distribution with n dimensional vector of parameters $(a_1, ..., a_n)$
$Exponential(\lambda)$	Exponential distribution with parameter λ
$Geo(p)$	Geometric distribution with parameter p
$Hypergeometric(n, h, r)$	Hypergeometric distribution with parameters n, h, r
$I(\cdot)$	Indicator function
$\mathcal{IG}(a, b)$	Inverse Gamma distribution with scale parameter a and shape parameter b
$i.i.d.$	Identically and independent distributed
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}_p(M, \Sigma)$	Multivariate Gaussian distribution with p-dimensional vector of means M and covariance matrix Σ
$\stackrel{\mathbb{P}}{\rightarrow}$	Convergence in law
$\mathcal{PG}(a, b)$	Pólya-Gamma distribution with parameters a and b
$Poisson(\lambda)$	Poisson distribution with parameter lambda
x^T	Transpose of vector x
$\mathcal{U}(a, b)$	Uniform distribution in (a, b)
$\mathcal{U}_d(a, b)$	Uniform discrete distribution in (a, b)

Introduction

In statistics, measurements of a set of variables that are repeated over time from the same unit are typically referred to as longitudinal data. Examples of longitudinal studies include tracking the academic performance of students over several years, observing the health condition of individuals at regular intervals to study disease progression, or understanding the economic trends of a country over several decades.

In biostatistics, the units of longitudinal studies are usually patients who are followed by researchers over an extended period, and their data are recorded at specific intervals. Those data can be used to track changes and patterns over time, making longitudinal studies particularly valuable for understanding trends, growth, and development, as well as investigating cause-and-effect relationships. In the following, we describe some advantages of the use of longitudinal data.

- Temporal Insight: longitudinal data allow researchers to examine changes and developments over time, capturing the dynamics of the variables under study.
- Individual Variation: by studying the same individuals over time, researchers can account for individual differences and analyze how different people respond to various factors.
- Causality: longitudinal studies can provide stronger evidence for establishing causal relationships between variables, as researchers can observe changes occurring before and after specific events or interventions.

However, many challenges can also arise with longitudinal data analyses. Some of the major ones are listed below.

- Censoring: some participants may drop out of the study over time, leading to missing data and potential bias.
- Time and Cost: conducting longitudinal studies can be time-consuming and expensive due to the extended data collection period.
- Compliance: participants repeated exposure to measurements can lead to changes in their behavior or responses, affecting the validity of the results.

Several statistical methods may be used to analyze longitudinal data, including repeated measures ANOVA, growth curve modeling, mixed-effects models, and many more. These techniques account for within-subject correlations and dependencies within the data, enabling more reliable conclusions regarding observed patterns and changes over time.

When analyzing longitudinal data, two primary statistical questions of interest arise, which lead to the application of different methodologies.

The first focus is on modeling, estimating, and forecasting the evolution of quantities of interest, for one or multiple individuals. This quantity could be the growth curve of a patient tumor or the concentration of a specific biomarker when examining clinical data, or the count of infected individuals during an epidemic.

The second focus is the analysis of repeated events over time, often referred to as recurrent events. These events typically pertain to adverse occurrences like migraines, heart attacks, respiratory obstructive events, epileptic seizures... In this scenario, the primary interest lies in studying the times between these recurrent events and their intensity over time. Researchers often compare these quantities between patients receiving an experimental treatment and patients receiving the standard of care to analyze the treatment efficacy.

This thesis presents three longitudinal studies based on real data. The first two studies pertain to the first class of problems, while the last one belongs to the second class of problems.

The first problem we address is the analysis of the Italian COVID-19 data by means of a new epidemic model. In March 2020, Italy, like the rest of the world, faced numerous challenges due to the COVID-19 pandemic. The scientific community dedicated several efforts to studying the SARS-CoV-2 virus from various angles, including medical, statistical, and economic perspectives. In particular, statisticians focused on modeling and forecasting the dynamics of the pandemic using data of different natures, including public daily data on infected and recovered patients.

Researchers applied classical tools in infectious disease modeling and statistical epidemiology [KM27; ISSb; WHO; Ita; AM92; Cap93; KR08; Mar15] and built upon the experience acquired during the previous SAR-CoV outbreak in 2002-2003 [Gum+04]. Many different models have been constructed, mostly using population dynamics or compartmental models. It became soon apparent that the well-known Susceptible-Infective-Recovered (SIR) model was too simplistic to describe the complexity of the pandemic [Mul21]. Consequently, numerous papers used extended versions of the SIR model to analyze the COVID-19 pandemic in Italy.

A widely recognized model is the SEPIA introduced in [Gat+20], which divides the population into nine categories: Susceptible, Exposed, Pre-symptomatic, Infected with symptoms, Asymptomatics, hospitalized, isolated, recovered, and deceased. Other relevan models are the SEIRD model [LZ20] (with Susceptibles, Exposed, Infected, Recovered, and Deceased), the SUIHTER model [Par+21] (with Susceptible, four different types of infected, namely Undetected with or without symptoms, Isolated, Hospitalized and Threatened, and finally Extinct and Recovered), the SIDARTHE model [Gio+20] (with Susceptible, Infected, Diagnosed, Ailing, Recognized, Threatened, Healed and Extinct), the SI^2R^2D model [BCV21] (with Susceptible, Infected not notified, Infected notified, Recovered not notified, Recovered notified, Deceased), and many others, e.g., [FP20]. For a comprehensive overview

of COVID-19 literature related to Italian data modeled with compartmental models, the reader can refer to [BL22].

Despite all the efforts, these models are hardly applicable to the analysis of public data. In most countries, the only data that are publicly available are the epidemic curves (the daily counts of the newly detected infections and that of the recovered and dead) and bed occupancies in the hospitals (both in ICU and other units). In this situation, inferring a complex model with numerous equations is impossible without making several arbitrary choices about the parameter values: data simply do not contain sufficient information to identify and estimate them all.

Our aim is to analyze the available data using a newly developed compartmental model called SIPRO (Susceptible-Infected-Positive-Recovered-Out model) that extends the well-known SIR model while maintaining simplicity and identifiability. Additionally, we seek to incorporate the heterogeneity of the Italian regions [Del+20], by combining different regional SIPRO models with a mixed-effect approach [Pra+20].

We exploit advanced Bayesian techniques to estimate the model parameters. Despite the abundance of available data, the complex nature of the pandemic makes inference challenging. Our primary focus is on estimating the asymptomatically infected individuals percentage and understanding the impact of social distancing measures on the pandemic progression in Italy. Furthermore, we estimate values during the first wave (i.e., February-June 2020) and compare our results with those obtained with the simple SIR model to assess the advantages and disadvantages of our new methodology. For our analysis, we use Italian public data collected by the Italian "Protezione Civile," available at the link [DPCa], starting from February 2020.

The second work is in collaboration with the researchers of the Urology department of the San Luigi Gonzaga Hospital, in Orbassano (Torino). In an observational study, they collected data from prostatectomized patients who underwent surgery due to prostate cancer and then had a clinical resurgence. As prostatectomized patients are at risk of resurgence, during a follow-up period, they have been monitored for prostate-specific-antigen (PSA) growth, an indicator of tumor progression.

The presence of tumor can be evaluated with an expensive exam, called Positron Emission Tomography with Prostate-Specific Membrane Antigen (PET-PSMA). To justify the high cost of the PET-PSMA and, at the same time, to contain the risk for the patient, this exam should be recommended only when there is strong evidence of tumor progression. Clinicians aim was estimating the optimal time to recommend the exam. There is a huge literature, both from a clinical and a statistical point of view, studying prostate cancer and, in particular, PSA, which has been proven to be one of the most significant biomarkers associated with this cancer [Sta+87; LUV08; Vic+09]. Tumor evolution and tumor resurgence have been directly modeled by ordinary differential equations models, as in [TT17], or through the PSA biomarker [Pea+91; Car+92; Mor+95; SC97; PT09; Hir+12].

With the development of new techniques and machinery to improve the early detection of the location of tumors and resurgence, the estimation of the optimal time to perform them has become a hot topic in the last decade. Since its introduction in 2016, PET-PSMA has gained high relevance. This technique has shown promising results in various clinical scenarios, including the initial staging of prostate cancer, detecting local recurrence after primary treatment, and identifying metastatic sites in men with rising PSA levels due to biochemical recurrence, as [Eib+16; Ver+16;

Afs+17; Fen+19; Fos+19; Hof+19; Reg+20] and many more. It can aid in the accurate assessment of the extent and location of prostate cancer, leading to better treatment planning and improved patient management. Some works also proved how the PET-PSMA examination results correlate with PSA [Per+19]. However, the estimation of the optimal time to perform this examination to early detect locations of disease is still an open problem. In [Lui+20], several biomarkers were studied to find the ones significantly associated with prostate resurgence that can be used as an indicator of the optimal time to perform the examinations.

The aim of our work is to construct a sensitive model to improve the estimation of time-to-examination. Our proposed model incorporates new information each time a new patient enters the dataset. Based on the patients history and collected data, we develop a hierarchical Bayesian model that jointly describes the PSA growth curve and the probability of a positive PET-PSMA result. By utilizing all past and present information about patients PSA measurements and PET-PSMA results, we aim to provide an informed estimate of the optimal time for examinations, thus enhancing current medical practices.

The third project is in collaboration with GSK Siena (Italy) and GSK Rixensart (Belgium) and aims at analyzing data from a clinical trial involving patients affected by Chronic Obstructive Pulmonary Disease (COPD). COPD is characterized by airflow obstructions known as acute exacerbations, categorized as mild, moderate, or severe. The frequency of these adverse events can vary significantly among patients. To address this issue, GSK developed an investigational vaccine that aims at reducing the frequency of moderate and severe exacerbations. Details about the safety and efficacy analysis conducted in phase 1 and phase 2b of the clinical trials can be found in [And+22; Aro+22].

Our aim is to re-analyze the clinical study data to make a sensitivity analysis, exploiting different models and making different assumptions about the missing data structure.

In the field of recurrent events, numerous models have been developed to understand the relationship between repeated occurrences of individual events. Examples of these models include the semi-parametric Andersen-Gill (AG) model [AG82] and the Prentice-Williams-Peterson model [PWP81]. Additionally, other approaches consider the interdependence of recurrent events by using shared random effects, as seen in joint frailty models [TD04], or by modeling it through a nuisance parameter, like the frailty parameter in a Negative Binomial model [Jah08].

However, the modeling task is further complicated by the presence of informative censoring, when individuals experiencing frequent exacerbations are more likely to withdraw from the study. In the literature, the Inverse Probability Censoring Weights approach (IPCW) [RR92; Rob93; RF00] is a well-known method to correct models while accounting for informative censoring. Nonetheless, corrections for informative censoring in the context of recurrent events, especially with time-varying covariates, are not commonly applied in practical scenarios [Wil+18].

The aim of our work is to provide a user-friendly introduction to IPCW for recurrent events data, explaining in detail how to apply the AG model, in conjunction with IPCW, based on the theoretical explanations given in [Mil+04]. We provide a step-by-step code to simulate and analyze the data. The method is finally applied to the COPD vaccine trial.

In addition, other collaborations with the *Laboratory for Cardiovascular Theranostics*, Cardiocentro Ticino Foundation (Lugano, Switzerland), and with the *Division of Internal Medicine and Hypertension*, Dipartimento di Medicina, Università degli Studi di Torino (Torino, Italy), produces to some medical papers not included in this thesis [Amo+20; Bur+20a; Bur+20b; Buf+21; Bur+21; Bur+22a; Bur+22b].

The thesis incorporates not only theoretical explanations and results but also the implemented and utilized `R` code (with `R` version $\geq$ 3.4). It is structured into two main parts. Part I presents evolution and dynamic problems and is composed of three chapters:

- in Chapter 1 we give some basics of Bayesian statistics and Markov Chain Monte Carlo algorithms;
- in Chapter 2 we present the analysis of the COVID-19 data;
- in Chapter 3 we present the analysis of the prostate cancer dataset.

Part II presents the recurrent events problem and is composed of three chapters:

- in Chapter 4 we provide some basics of survival analysis;
- in Chapter 5 we provide some basics of recurrent events settings and algorithms;
- in Chapter 6 we present the analysis of the COPD dataset.

Finally, in Chapter 7, we sum up the results and give some overall considerations, discussing future research lines.

Part I

Bayesian inference for the dynamics of longitudinal data

Chapter 1

Background of Markov Chain Monte Carlo methods

This section shortly gives the fundamentals of Bayesian parametric inference which the first part of the thesis is based on. In particular, we give some fundamentals of Markov Chain Monte Carlo (MCMC) algorithms and related techniques used and learned before and during the Ph.D. period. No full details are given, but some references are provided.

We introduce the concept of prior, posterior, proposal, and predictive densities, and we clarify how these quantities are related to the standard Gibbs update. Then, we go into detail about how to define conjugate prior in easy and more complex settings, giving details on the Metropolis-Hastings algorithm and discussing how to tune and adapt the proposal densities. Finally, we describe the parallel tempering algorithm to deal with multimodal problems. In the following, we refer to random variables with capital letters and to their realizations with small ones.

The general application of MCMC algorithms is to infer the distributions of the parameters of a model of interest using the information given by the collected data. We call $\{y_i\}_{i=1}^n$ the data, which are independently collected from the distribution $f_{Y\,|\,\Theta}(\,\cdot\,|\,\theta)$ with unknown p-dimensional parameter Θ. The main statistical interest is estimating θ and its conditional distribution $f_{\Theta|\,Y}(\,\cdot\,|\,y_1,\ldots,y_n)$. The main idea of MCMC methods is to infer the latter distribution using Bayes' theorem, namely

$$f_{\Theta|Y}(\theta\,|\,y_1,\ldots,y_n) = \frac{f_{Y|\Theta}(y_1,\ldots,y_n\,|\,\theta) f_\Theta(\theta)}{f_Y(y_1,\ldots,y_n)}. \tag{1.1}$$

In particular, MCMC methods aim at constructing a Markov Chain whose stationary distribution is the target density $f_{\Theta\,|\,Y}$. Thus, simulating realizations of the chain for a sufficient number of iterations provides samples from the stationary distribution of interest. Generally, statisticians refer to $f_\Theta(\,\cdot\,)$ as prior density, as it encapsulates all the prior information available about the parameter of interest, to $f_{Y|\Theta}(\,\cdot\,|\,\theta)$ as likelihood, and to the target density $f_{\Theta|Y}(\,\cdot\,|\,y_1,\ldots,y_n)$ as posterior density.

1.1 Gibbs sampler

The problem of estimating the p-dimensional density of Θ is usually decomposed in p smaller steps, thanks to the Gibbs Sampler scheme (GS). This iterative scheme is particularly useful when we can simulate from the single conditional densities $f_{\Theta_j|\Theta_{(-j)}}(\,\cdot\,|\,y_1,\ldots,y_n,\theta_{(-j)})$, also called full conditionals, for each component j of Θ, where $\Theta_{(-j)}$ is the vector Θ except for the j-th component, and $\theta_{(-j)}$ its realization.

After initializing the vector Θ of parameters to θ^0, the b-th iteration of the Gibbs sampler can be summed up as follow:

- step 1: simulate Θ_1^b from $f_{\Theta_1|\Theta_{(-1)}}(\,\cdot\,|\,y_1,\ldots,y_n,\theta_2^{b-1},\theta_3^{b-1},\ldots,\theta_p^{b-1})$,
- step 2: simulate Θ_2^b from $f_{\Theta_2|\Theta_{(-2)}}(\,\cdot\,|\,y_1,\ldots,y_n,\theta_1^{b},\theta_3^{b-1},\ldots,\theta_p^{b-1})$,

  $\vdots$
- step p: simulate Θ_p^b from $f_{\Theta_p|\Theta_{(-p)}}(\,\cdot\,|\,y_1,\ldots,y_n,\theta_1^{b},\theta_2^{b},\ldots,\theta_{p-1}^{b})$.

This method requires a good knowledge of the conditional densities: for particular scenarios, some choices of prior densities $f_\Theta(\,\cdot\,)$, called conjugate densities, allow to directly compute the conditional densities of interest and to exploit the Gibbs sampler structure. In the following, we report two particular scenarios that are widely used later on in the thesis, for normal and binomial data problems.

The first case we present is the Gaussian-likelihood problem: if $\{y_i\}_{i=1}^n$ are Gaussian distributed data (e.g., the temperature measured in some buildings) the marginal density is then

$$y_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu,\sigma^2), \tag{1.2}$$

where $\Theta = (\mu,\sigma)$ is the 2-dimensional parameter of interest. Assuming priors $\mu \sim \mathcal{N}(\mu_0,\sigma_0^2)$ and $\sigma^2 \sim \mathcal{IG}(a_1,a_2)$, the Gibbs scheme iteratively samples, for each iteration b, from the following distributions

$$\begin{aligned} \mu^b\,|\,y_i,\sigma^{2^{b-1}} &\sim \mathcal{N}\left(\frac{1}{1/\sigma_0^2+n/\sigma^{2^{b-1}}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n y_i}{\sigma^{2^{b-1}}}\right),\frac{\sigma_0^2+\sigma^{2^{b-1}}}{\sigma^{2^{b-1}}+n\sigma_0^2}\right),\\ \sigma^{2^b}\,|\,y_i,\mu^b &\sim \mathcal{IG}\left(a_1+\frac{n}{2},a_2+\frac{\sum_{i=1}^n(y_i-\mu^b)^2}{2}\right). \end{aligned} \tag{1.3}$$

The example can be extended to include baseline covariates $Z_i = (z_{1i},\ldots,z_{pi})^T$ collected for each observation (e.g., the architectural features of each building under analysis), to obtain the new model

$$y_i \overset{i.i.d.}{\sim} \mathcal{N}(Z_i^T\mu,\sigma^2),\ \text{with}\ \mu = (\mu_1\ldots\mu_p)^T. \tag{1.4}$$

In this case, the vector of parameters is $\Theta = (\mu_1,\ldots\mu_p,\sigma)$. A good prior choice is to assume $\mu \sim \mathcal{N}_p(M,V)$ and $\sigma^2 \sim \mathcal{IG}(a_1,a_2)$ and to iterate the GSS scheme

$$\begin{aligned} \mu^b\,|\,y_i,\sigma^{2^{b-1}} &\sim \mathcal{N}_p(M_p^{b-1},V_p^{b-1}),\\ \sigma^{2^b}\,|\,y_i,\mu^b &\sim \mathcal{IG}\left(a_1+\frac{n}{2},a_2+\frac{\sum_{i=1}^n(y_i-Z_i^T\mu^b)^2}{2}\right). \end{aligned} \tag{1.5}$$

where $V_p^{b-1} = \left(\frac{1}{\sigma^{2,b-1}}Z^TZ+V^{-1}\right)^{-1}$ and $M_p^{b-1} = V_p^{b-1}\left(\frac{1}{\sigma^{2,b-1}}Z_i^Ty+V^{-1}M\right)$, with $Z = (Z_1,...,Z_n)$ matrix of the data of dimension $p\times n$ and y column vector of the data.

The other common scenario we describe is the binomial one. Let $\{y_i\}_{i=1}^n$ be the number of successes obtained out of n_i independent trials (e.g., the number of positive/negative examinations a patient had over a fixed period of time) and $Z_i = (z_{1i},\ldots,z_{pi})^T$ the vector of measured covariates collected for each observation (e.g., physical characteristics of each patient), then the marginal density of the model

is

$$y_i \overset{i.i.d.}{\sim} Bin\left(n_i, \frac{1}{1+\exp(-Z_i^T\mu)}\right). \tag{1.6}$$

Polson et al. [PSW13] suggest using a Gaussian prior $\mu \sim \mathcal{N}_p(\mu_0, V)$, and exploit the Pólya-Gamma Gibbs update which gives the following full conditionals for the parameters μ and the latent variables ω_i, which are needed for the algorithm

$$\begin{aligned} \omega_i^b \mid \mu^{b-1} &\sim \mathcal{PG}(n_i, Z_i^T\mu^{b-1}), \\ \mu^b \mid y_i, \omega^b &\sim \mathcal{N}(m_{\omega^b}, V_{\omega^b}), \end{aligned} \tag{1.7}$$

where

$$\begin{aligned} V_{\omega^b} &= (Z^T\Omega^b Z + V^{-1})^{-1}, \\ m_{\omega^b} &= V_{\omega^b}(Z^T k + V^{-1}\mu_0^T), \end{aligned} \tag{1.8}$$

with $k = (y_1 - n_1/2, \ldots, y_n - n_n/2)$, $\Omega^b = diag(\omega_1^b, \ldots, \omega_n^b)$, and Z the matrix of the full covariate, defined as for the Gaussian case.

This update derives from the intrinsic structure of the Pólya-Gamma density:

Definition 1.1.1 (Pólya-Gamma distribution)**.** *A random variable ω has a Pólya-Gamma distrubution with parameters $d > 0$ and $c \in \mathbb{R}$, indicated with $\omega \sim \mathcal{PG}(d, c)$, if*

$$\omega \overset{D}{=} \frac{1}{2\pi^2}\sum_{h=1}^{\infty}\frac{G_k}{(k-1/2)^2 + c^2/(4\pi^2)}, \quad \textit{where } G_k \overset{i.i.d.}{\sim} \Gamma(d, 1). \tag{1.9}$$

Polson et al. [PSW13] show that the binomial likelihood parametrized by log-odds can be represented as a mixture of Gaussians variables with respect to the Pólya-Gamma distribution. In particular, they proved the Theorem 1.1.1.

Theorem 1.1.1 (Theorem Pólya-Gamma)**.** *Let $p_{d,0}(\omega)$ denote the density of a random variable $\omega \sim \mathcal{PG}(d, 0)$, with $d > 0$, then the following integral identity holds for all $a \in \mathbb{R}$*

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^d} = 2^{-d}e^{k\psi}\int_0^{\infty} e^{-\omega\psi^2/2}p_{d,0}(\omega)d\omega, \tag{1.10}$$

with $k = a - d/2$. Moreover, the conditional distribution of $\omega \mid \psi$ is $p(\omega \mid \psi) \sim \mathcal{PG}(d, \psi)$, with ψ linear function of predictors.

Definition 1.1.1 and Theorem 1.1.1 can be exploited to write the single observation contribution to the likelihood, where $\psi_i = Z_i^T\mu$, as

$$L_i(\mu) = \frac{(\exp(Z_i^T\mu))^{y_i}}{1+\exp(Z_i^T\mu)} \propto \exp(k_i Z_i^T\mu)\int_0^{\infty}\exp(-\omega_i(Z_i^T\mu)^2/2)p_{n_i,0}(\omega_i)d\omega_i, \tag{1.11}$$

with $k_i = y_i - n_i/2$. Extending Equation (1.11) to the product over all the observation leads to the conditional density

$$\begin{aligned} p(\mu \mid \omega_1, .., \omega_n, y_1, ..., y_n) &\propto p(\mu) \prod_{i=1}^{n} L_i(\mu|\omega_i) \\ &= p(\mu) \prod_{i=1}^{n} exp(k_i Z_i^T \mu - \omega_i (Z_i^T \mu)^2/2) \\ &\propto p(\mu) \prod_{i=1}^{n} exp\left(\frac{\omega_i}{2} (Z_i^T \mu - k_i/\omega_i)^2 \right) \\ &\propto p(\mu) exp\left(-\frac{1}{2} (m - Z\mu)^T \Omega (m - Z\mu) \right), \end{aligned} \tag{1.12}$$

with $m = (k_1/\omega_1, ...k_n/\omega_n)$ and $\Omega = diag(\omega_1, ..., \omega_n)$. It can easily be seen that a Gaussian prior on μ leads to the Gibbs scheme in Equation (1.7).

1.1.1 Metropolis-Hastings algorithm

Whenever the conditional densities are not directly computable, some other schemes should be used to sample. The Metropolis-Hastings algorithm (MH) is an iterative procedure to sample from a general target distribution $f_{\Theta|Y}$: starting from an appropriate value θ^0 for Θ, for each iteration b, it proposes a new value θ^* which can be properly accepted or rejected. The current value of Θ at the beginning of iteration b is here denoted by θ^{b-1}. The single update scheme, for iteration b, is the following

- step 1: propose θ^* from the distribution $Q(\theta)$ with density function $q(\cdot \mid \theta^{b-1})$;
- step 2: set

$$\theta^b = \begin{cases} \theta^* & \text{with probability } a(\theta^{b-1}, \theta^*), \\ \theta^{b-1} & \text{otherwise,} \end{cases} \tag{1.13}$$

  where

$$a(\theta^{b-1}, \theta^*) = \frac{f_{\Theta|Y}(\theta^* \mid y_1, \ldots, y_n) q(\theta^{b-1} \mid \theta^*)}{f_{\Theta|Y}(\theta^{b-1} \mid y_1, \ldots, y_n) q(\theta^* \mid \theta^{b-1})} \tag{1.14}$$

  is called acceptance ratio and is usually difficult to compute, we refer to it as a^b for brevity. Exploiting Bayes theorem this can be rewritten as

$$a(\theta^{b-1}, \theta^*) = \frac{f_{Y|\Theta}(y_1, \ldots, y_n \mid \theta^*) f_\Theta(\theta^*) q(\theta^{b-1} \mid \theta^*)}{f_{Y|\Theta}(y_1, \ldots, y_n \mid \theta^{b-1}) f_\Theta(\theta^{b-1}) q(\theta^* \mid \theta^{b-1})}, \tag{1.15}$$

  which is easier to compute.

In particular, we refer to $q(\cdot)$ as the proposal density, as it is used to sample proposal candidates. The prior and proposal densities, initial conditions, and the number of iterations are user-defined values that significantly impact the algorithm's performance. Selecting an adequate number of iterations is crucial to ensure convergence. Furthermore, the initial conditions have a substantial influence on the first few iterations; thus, it is recommended to discard a certain number of samples. This step is called Burn-in. Additionally, if the values exhibit high correlation, it is recommended to retain only every k-th sample instead of all of them. This process is referred to as *thinning*. Finally, by exploiting the obtained estimation,

we can make predictions for the new variable $\tilde{Y}$ through the predictive distribution $f_{\tilde{Y}|Y}(\cdot \mid y_1, \ldots, y_n)$.

Choosing the proposal density is the most challenging part of the MH algorithm. One of the most exploited proposal densities is based on random walk structure, in particular on the symmetric normal random walk for parameters defined on $\mathbb{R}$. The main advantage of symmetrical proposals is that they cancel out from the acceptance ratio a^b. If Θ is a one-dimensional parameter, the proposal density for θ^*, at iteration b, is equal to

$$q(\theta^* \mid \theta^{b-1}) = \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2\omega^2}(\theta^* - \theta^{b-1})^2\right), \tag{1.16}$$

where the standard deviation ω of the random walk step is a user-defined parameter that has a big impact on the rate of convergence and on the final results, and it is not trivial to be defined. In particular, ideally, the aim is to initially use ω big enough to explore all the admitted values for Θ, and, once the algorithm reaches convergence, to use a smaller ω to better explore the posterior. The best solution is to iteratively adapt this quantity as part of the algorithm, allowing the variance to enlarge or decrease if the acceptance rate is not close to a user-defined threshold (ideally the ratio at iteration b should be $a^b \simeq 0.25$), as explained in [AT08]. We present here some strategies to automatically tune ω in the univariate and the multivariate case.

For the univariate case, the adaptive proposal variance should be updated as

$$\log(\omega^b) = \log(\omega^{b-1}) + \gamma^b(\hat{a}^{b-1} - a^*), \tag{1.17}$$

where $\{\gamma^b\} \subset (0, +\infty)^{\mathbb{N}}$ is a user-defined sequence of possible stepsize, $\hat{a}^{b-1}$ is the acceptance ratio evaluated at iteration $b-1$ and a^* the desired acceptance ratio (usually around 0.25) to achieve through the adaptive method. The non-increasing sequence $\{\gamma^b\}$ satisfies the low decay assumption [AM06] such that $\sum_{b=1}^{\infty} \gamma^b = \infty$, and $\sum_{b=1}^{\infty} (\gamma^b)^{1+\delta} < \infty$ for some $\delta < 0$. The full detailed theory about how to properly tune θ can be found in [AT08]. The update step can be performed at each iteration, but sometimes it is preferred to do it every m iteration to give the MCMC procedure some steps to stabilize before adapting the proposal variance. In this case

$$\hat{a}^{b-1} = \frac{\sum_{k=b-m}^{b-1} a^k}{m}. \tag{1.18}$$

The multivariate problem is an extension of the univariate case. If Θ is a p-dimensional parameter, the proposal density for θ^*, at iteration j, is

$$q(\theta^* \mid \theta^{b-1}) = \frac{1}{\sqrt{\det\Omega}(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(\theta^* - \theta^{b-1})^T \Omega^{-1} (\theta^* - \theta^{b-1})\right), \tag{1.19}$$

where Ω is the covariance matrix that we need to carefully define. The adaptive procedure (procedure 4 in [AT08]) suggests, at each iteration b, modifying Equation (1.19) as

$$q(\theta^* | \theta^{b-1}) = \frac{1}{\sqrt{\det(\lambda^{b-1}\Omega^{b-1})}(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(\theta^* - \theta^{b-1})^T (\lambda^{b-1}\Omega^{b-1})^{-1} (\theta^* - \theta^b)\right), \tag{1.20}$$

where

$$\begin{aligned} \log(\lambda^b) &= \log \lambda^{b-1} + \gamma^b[a^{b-1} - a^*], \\ \mu^b &= \mu^{b-1} + \gamma^b(\theta^b - \mu^{b-1}), \\ \Omega^b &= \Omega^{b-1} + \gamma^b[(\theta^b - \mu^{b-1})(\theta^b - \mu^{b-1})^T - \Omega^{b-1}]. \end{aligned} \tag{1.21}$$

The quantities θ^0, μ^0, Ω^0, and the non-increasing sequence $\{\gamma^b\}$ (see details above) are user-defined.

Sometimes, the MH and GS algorithms are utilized in combination to estimate a high-dimensional vector of parameters: the presented MH algorithm can be used to sample from the target density $f_{\Theta_j|\Theta_{(-j)}}(\cdot \,|\, y_1, \ldots, y_n, \theta_{(-j)})$ and then included into the GS steps. We refer to this procedure as the Metropolis-within-Gibbs update. In such cases, certain parameters are simulated by conditional distributions, while others require a proposal via an MH update. We refer to [ST10] for further details about Gibbs Sampler. While, for the MH algorithm, we refer to the first published paper in which this algorithm was proposed [Met+53] and to a more recent one [RC10].

1.2 Bayesian cross-validation predictive density

Up to this point, we have defined prior, posterior, proposal, and predictive densities as the fundamental components of Bayesian inference. Yet, another density of significant statistical interest can be introduced: the cross-validation predictive density, a tool to assess the accuracy and reliability of predictive models. This density is instrumental in evaluating the generalization performance of statistical models by systematically partitioning the dataset into training and validation sets. Through this iterative process, the model predictive power is tested on data not used during the training phase, allowing for a robust assessment of its predictive accuracy. This methodology is particularly useful when overfitting or underfitting may pose challenges to model generalization. The predictive density provides a representation of the uncertainty associated with the model predictions, helping researchers to make informed decisions based on a deep understanding of the model reliability across different data scenarios. In this Section, we define the cross-validation predictive density and give small insights into its theoretical justification and practical implementation. We refer to [GRS95] for additional details.

For the set of observation $\mathbf{y} = \{y_i\}_{i=1}^n$, the cross-validation predictive densities is defined to be the set

$$\{f_{Y_i|Y_{(-i)}}(\cdot|\mathbf{y}_{(-i)}), i = 1, 2, ..., n\},$$

where $\mathbf{y}_{(-i)}$ denotes the set of all elements of $\mathbf{y}$ except y_i, and $Y_{(-i)}$ and Y_i are respectively the random variables corresponding to $\mathbf{y}_{(-\mathbf{i})}$ and y_i.
Cross-validation predictive densities are usually calculated as

$$f_{Y_i|Y_{(-i)}}(y_i|\mathbf{y}_{(-i)}) = \int f_{Y_i|\Theta,Y_{(-i)}}(y_i|\theta, \mathbf{y}_{(-i)}) \cdot f_{\Theta|Y_{(-i)}}(\theta|\mathbf{y}_{(-i)})d\theta.$$

The distributions $f_{Y_i|Y_{(i)}}(y_i|\mathbf{y}_{(i)})$ permit us to work with univariate distributions. Single-element deletion is standard in classical regression diagnostics and is called leave-one-out cross-validation. However, we can also work with set deletion rather than a single element, and the presented theory can be extended to this latter case.

For clarity, we omit the subscript from distributions in what follows (in particular, in Equations 1.22-1.23-1.24). From a computational point of view, the cross-validation predictive density can be written as

$$\begin{aligned} f(y_i|\mathbf{y}_{(-i)}) &= \frac{f(\mathbf{y})}{f(\mathbf{y}_{(-i)})} \\ &= \left(\int \frac{f(\mathbf{y}_{(-i)}, \theta)}{f(\mathbf{y}, \theta)} f(\theta|\mathbf{y}) d\theta \right)^{-1} \\ &= \left(\int \frac{1}{f(y_i|\mathbf{y}_{(-i)}, \theta)} f(\theta|\mathbf{y}) d\theta \right)^{-1}, \end{aligned} \tag{1.22}$$

that can be approximated, through Monte Carlo integration, as

$$\hat{f}(y_i|\mathbf{y}_{(-i)}) = \frac{1}{\frac{1}{B}\sum_{b=1}^{B} \frac{1}{f(y_i|\mathbf{y}_{(-i)}, \theta^b)}}, \tag{1.23}$$

where $\{\theta^b,\ b = 1, ..., B\}$ is the posterior sample for Θ. Equation (1.23) can be utilized to sample from the cross-validation density of interest in an importance-sampling scheme. The key advantage of this method is that there is no necessity to rerun the MCMC sampler using only $\mathbf{y}_{(-i)}$ to obtain samples from the cross-validation density $f(y_i|\mathbf{y}_{(-i)})$. In fact, by appropriately leveraging the set $\{\theta^b,\ b = 1, ..., B\}$ we can obtain the sample $\{\theta^{b*},\ b = 1, ..., B\}$ from $f(\theta|\mathbf{y}_{(-i)})$ which enable us to draw $\{\mathbf{y}^{b*}\}$ from $f(\mathbf{y}|\theta^{b*})$, resulting in its component y_i^{b*} beeing a sample $f(y_i|\mathbf{y}_{(-i)})$. Thus, to sample θ^{b*}, we only need to compute the importance ratio

$$\frac{f(\theta^b|\mathbf{y}_{(-i)})}{f(\theta^b|\mathbf{y})} \propto \frac{1}{f(y_i|\mathbf{y}_{(-i)}, \theta^b)} = w^b, \tag{1.24}$$

and subsequently resample from the set $\{\theta^b\}$ with probabilities proportional to $\{w^b\}$. Further details on how to utilize the importance sampling scheme to sample from cross-validation densities can be found in Section 9.6.3 of [GRS95].

1.3 Parallel Tempering

Estimating the parameters of complex hierarchical models is the target of MCMC algorithms. Even when combining proper MH and GS updates, priors, and proposals, problems characterized by multimodal density functions make the inference very challenging. Specifically, the standard MCMC techniques described earlier may fail or have a long convergence time, as the simulated parameter values can get trapped in incorrect high-density regions of the parameter space.

To address complex density functions and low-identifiability problems, Sambridge [Sam13] proposed a meta-algorithm that can be combined with MCMC updates to improve the exploration of the parameter space, called Parallel Tempering (PT). The main idea behind this technique is to run some MCMC methods, each with likelihood proportional to $f_{Y|\Theta}^{1/T_l}(y_1, \ldots, y_n \mid \theta)$, each differing for the temperature parameter $\{T_l\}_{l=1,...,L}$. The method then proposes swap-updates to exchange the values of parameters between different chains. The role of the temperature parameter is to flatten the multimodal distribution, allowing the chains to explore the space more

effectively. At iteration b, $\theta^{b,j}$ is the sampled values for chain j with temperature T_j and $\theta^{b,k}$ the sampled values for chain k with temperature T_k. The main idea is to perform an MH step with the deterministic proposal to swap the parameters between the chains. The corresponding acceptance ratio at iteration b, for chain j and k is

$$a^b(j,k) = \min\left(1, \left[\frac{f_{Y|\Theta}(y_1,\ldots,y_n \mid \theta^{b,k}) f_\Theta(\theta^{b,k})}{f_{Y|\Theta}(y_1,\ldots,y_n \mid \theta^{b,j}) f_\Theta(\theta^{b,j})}\right]^{\frac{1}{T_j}} \left[\frac{f_{Y|\Theta}(y1,\ldots,y_n \mid \theta^{b,j}) f_\Theta(\theta^{b,j})}{f_{Y|\Theta}(y_1,\ldots,y_n \mid \theta^{b,k}) f_\Theta(\theta^{b,k})}\right]^{\frac{1}{T_k}}\right). \tag{1.25}$$

The vector of temperatures $T = (T_1 = 1 < T_2 < T_3 < \cdots < T_L)$ should be carefully defined by the user. The swap update is usually performed between couples of neighboring temperatures. The base PT update should not be done at each iteration: usually, a single PT update is performed after a batch of standard MCMC iterations is run (for each temperature). The procedure is then repeated multiple times and the results associated with $T_l = 1$ are retained to form the desired sample.

Chapter 2

Analyzing the COVID-19 pandemic in Italy with the SIPRO model

Part of this chapter has been published as the following conference communication: Amongero, Martina and Bibbona, Enrico and Mastrantonio, Gianluca (2021) **Analyzing the COVID-19 pandemic in Italy with the SIPRO model.** In: Book of short papers - SIS 2021, pp. 1568–1573. ISBN: 9788891927361. [ABM21]

Understanding disease transmission and modeling the spread of infectious diseases has attracted a lot of interest from mathematicians and statisticians. In 2019, with the worldwide COVID-19 pandemic, epidemic modeling has become a hot topic. Researchers focused on very different aspects of the problem, using different data: household transmissions, generation intervals, the spread of the different variants, etc. These problems motivated the development of new approaches, models, techniques, and algorithms.

In this Chapter, we focus on the formulation of a new mathematical model that incorporates the presence of unaccounted infections, estimated as latent quantities. The model can be applied to analyze the publicly available data, to model and forecast the spread of the epidemic, and the effectiveness of public policies.

To compare the new approach with the existing literature, we shortly recall the definitions of the main quantities that are usually monitored during an epidemic, such as the basic reproduction number, and then we review the two main modeling approaches that have been proposed: mechanistic models and phenomenological ones. Mechanistic models try to interpret the data by identifying the underlying basic mechanisms. Phenomenological models instead focus on making accurate forecasts, no matter what the data-generating mechanism is.

Then, we focus on compartmental models, introducing the well-known Susceptible-Infected-Recovered model (SIR). Our new model, the Susceptible-Infected-Positive-Recovered-Out (SIPRO) model, is then explained in detail, and applied to model the time course of the epidemic in each of the 21 Italian regions. To set up our inferential procedure, we integrate our SIPRO model in a statistical framework where the distribution of the measurement errors is fully specified together with a mixed effect model that takes into account regional heterogeneities. We perform a simulation study to investigate the strengths and weaknesses of our proposal and then apply it to the real data collected by the Italian Protezione Civile [DPCa][1].

2.1 Background of Epidemics models for COVID-19 data

All over the world, the COVID-19 pandemic has had a terrible impact on everyday life, causing nearly 6.486.277 deaths by the end of July 2023. The first cases of COVID-19 were described in December 2019 in Hubei, China. In a few months, the virus spread worldwide, and on March 11, 2020, the World Health Organization declared the pandemic stage. Given the severity of the disease, COVID-19 has gained the attention of many researchers, who studied its evolution from different points of view in the last two years. Many statisticians are contributing in different ways. We discuss here the interest in mathematical models (mechanistic of phenomenological) and the related inference task. Such models can help, in particular, the surveillance of the epidemic, the evaluation of public policies, and the estimation of vaccine efficacy. In this direction, the most interesting quantities are:

- number of people who get infected $I(t)$ over time (which is usually measured in days);
- the mean number of new cases deriving from a single infectious one (both the so-called *basic reproduction number* ρ^0 and *effective reproduction number* $\rho^{\text{eff}}(t)$), defined below;
- the generation time s: the difference between the time of infection of an individual and the time of infection of one of the people the individual got in contact with. Usually, the probability of getting infected after day s is indicated with ω_s.

These quantities are generally linked by the renewal equations:

$$I(t) = \rho^{\text{eff}}(t) \sum_{s=1}^{t} I(t-s)\omega_s, \tag{2.1}$$

where s is the number of infectiveness days. For further explanation, see [Fra07; AMM22].

Definition 2.1.1. *The basic reproduction number is a baseline quantity and gives the mean number of new infections from an individual in the population who happens to be the first and only one infected.*

Definition 2.1.2. *The effective reproduction number is a time-dependent quantity, which gives the mean number of new infections from each infective individual in the population at the current state at time t (when other individuals are possibly infected).*

Values of the effective reproduction number greater than one indicate that the epidemic will spread, while values smaller than one indicate that the pandemic will eventually end. This is the reason why the government decisions were mostly based on this quantity and its estimation became the goal of many approaches developed during the last years.

Many methods have been proposed to estimate $\rho^{\text{eff}}(t)$, but all of them need $I(t)$ to be known (or estimated). Thus, the biggest difficulty in estimating the effective reproduction number is to make inference on $I(t)$ despite some data are not available: the only data that can be collected are the number of positive people, the number of hospitalized people, the number of people with complications, and the number of deaths. On the other hand, the number of asymptomatic people (who do not

make a nasopharyngeal swab) is unknown, as is the number of recovered people who were not tested. However, the asymptomatic part of the population makes the most significant contribution to the pandemic spread.

In Italy, the first ascertained case of COVID-19 was detected on January 30, 2020, when two Chinese tourists tested positive. On February 21, the first outbreak was found in Lombardy, starting a very quick growth in the number of cases that anticipated the trend that was later seen in most European Countries and almost everywhere in the world. Since February 24, Italy started to keep track of the daily counts of people that tested positive, and of those that have been declared removed in each of the 21 Italian administrative divisions called *regions*. Official data are publicly available at [DPCa]. Most of the literature we briefly recap here takes advantage of these data. Few papers were able to also include data from the asymptomatic part of the population, one of the most interesting analyses on Italian data being [Lav+20]. This study is based on the population of Vo', an Italian municipality. Here more than 75% of the population was tested, even if without symptoms. However, note that this work uses a restricted dataset which leads to very wide confidence intervals of the parameters of interest.

2.1.1 Phenomenological models

Phenomenological models manage to give excellent estimations and forecasts at the cost of losing the interpretability of the model describing the data, allowing for the construction of more flexible models. In addition, modeling directly the quantity of interest, without having to construct the whole structure to mimic the whole dynamics, speeds up computational and inference times. Many different structures and models were used to analyze the Italian public data.

In [Ala+21], the authors proposed a parametric regression model based on the use of the Richards' curve (a generalized logistic function) in place of the widely used exponential or polynomial trends. In particular, they replace the Gaussian assumption for the distribution of log-daily counts (positive, infected,...) with the Poisson and/or Negative Binomial distributions for counts. The peculiarity of Richards' curve lies in the ability to describe a great variety of growing processes, which includes as special cases the standard logistic growth curve and the Gompertz growth curve. On the other hand, Bonifazi et al. [Bon+21] studied Equation (2.1) using an exponential function to estimate the number of infected people and a fixed mean generation time, computed by the Italian Istituto Superiore di Sanità. Those are just two examples of the works published in this direction, but many other ones could be mentioned.

However, as the main interest of our contribution is to analyze the applicability of compartment models more than phenomenological ones, we do not make a full overview of this literature but, instead, we focus more on the procedure utilized by the Italian government, used later on in this chapter. The procedure applied by the government to analyze the pandemic evolution and to make informed decisions is based on a Poisson distribution assumption as

$$\begin{aligned} C(t) - I_s(t) &\overset{i.i.d.}{\sim} Pois\Big(G(t)\sum_s \omega(s)C(t-s)\Big), \\ \rho^{\text{eff}}(t) &= \frac{1}{7}\sum_{s=0}^{6} G(t-s), \end{aligned} \tag{2.2}$$

where $C(t)$ is the number of symptomatic cases with symptoms arising on day t, $G(t)$ is the daily transmissibility, $I_s(t)$ is the number of symptomatic cases with symptoms arising on day t and arriving from a different region, and ω_s defined as before. The big advantage of this procedure was the possibility to exploit the information contained in more detailed data owned by the Italian Government [GM20].

2.1.2 Compartmental models

Mechanistic models are highly popular as they interpret the data based on the underlying mechanism. Compartmental models are a very general tool often applied to describe the evolution of an infectious disease: a systematic review, published in 2021 [Gna+21], revealed that 46.1% of the proposed models so far was based on compartmental approaches.

The population is divided into labeled compartments (e.g., susceptible, symptomatic, recovered, and so on). Each individual is originally assigned to one of these compartments and can move between compartments at some rates. The compartmental structure mimics the real word spread of the infection. The main focus is studying nature, time, and rate of flows between compartments. The most famous and simplest compartmental models are the SIR and the Susceptible-Infectious-Susceptible (SIS) [KM27].

For the intrinsic structure of the COVID-19 pandemic and the available measured data, the direct application of these basic and well-known models was possible but not realistic. The reason being that only symptomatic people were tested and counted in the daily number of infected cases, while the asymptomatic part of the population, which was mainly responsible for new cases, was not measured [Gae20; Mul21]. Consequently, applying the SIR or SIS model gives a wrong interpretation of reality. To address this issue, many modifications of the SIR model have been proposed, accounting for new compartments: asymptomatic, hospitalized, Intensive Care Unit beds (ICU), deaths, and many more. For instance, the Italian public data were analyzed in [Gio+20] with a SIDARTHE model, which accounts for symptomatic and asymptomatic cases, both detected or not. Similarly, the SEIRD model proposed in [LZ20] accounts also for the exposed people. The SEPIA model [Gat+20] divides the population into nine categories: Susceptible, Exposed, Pre-symptomatic, Infected with symptoms, Asymptomatics, hospitalized, isolated, recovered, and deceased. Another model is the SUIHTER [Par+21] that considers Susceptible, and four different types of infected, namely Undetected with or without symptoms, Isolated, Hospitalized and Threatened, and finally Extinct and Recovered. Some models strongly rely on daily dead count: in [BCV21], the authors define a compartmental model SI^2R^2D that accounts for both detected and undetected infections and assumes that only notified cases can die (compartments are: Susceptible, Infected not notified, Infected notified, Recovered not notified, Recovered notified and Deceased). Finally, many other models, such as the one in [Gat+20], also include the migration phenomena.

To compare and summarize the vast number of proposed compartmental models, several meta-analyses, comparative papers, and reviews are available [CVB22; Kon+22]. For a comprehensive overview of COVID-19 literature related to Italian data modeled with compartmental models, we refer the reader to [BL22]. Furthermore, as the pandemic evolved, new data were collected, and new tasks needed to be solved: with the introduction of vaccines, in December 2020, all compartment models needed updates to include the immune segment of the population. For example,

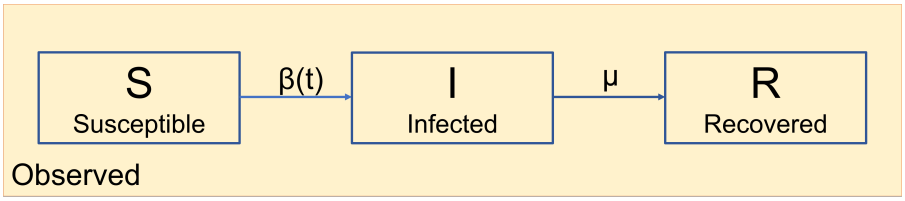

FIGURE 2.1: Graphical representation of SIR model.

the SIDARTHE model was modified to include vaccinations in [Gio+21]; moreover, over time, the SARS-CoV-2 virus also mutated, requiring new models to distinguish between different variants [Cal+21; FRL22; Del+23].

SIR model

The most famous compartmental model for epidemic dynamics is the SIR model [KM27]. It divides a population of m individuals into three groups: people who never had the illness before and can be infected by people who are infectious, respectively S and I, and finally, people who recover from the infection and become immune or died, R (see Figure 2.1).

Several simplifying assumptions are made: the population is assumed to be closed, homogeneous, and homogeneously mixed. Moreover, the latent periods, the time-varying infectivity, and the practical immunity are not taken into account. We refer to the transition parameter associated with the infection (going from S to I) as β and to the recovery parameter, that regulates the transition from I to R, as μ.

It also exists a stochastic version of the SIR model, that can be expressed through the chemical reaction network reported in the reactions table below (Table 2.1), involving three species $\{S, I, R\}$, namely Susceptibles, Infected, and Recovered, two reactions $\{r_1, r_2\}$, one that regulates the infections and one that regulates the recovery, and four complexes $\{S + I, 2I, R, I\}$ [AK15].

The stochastic model assumes the infectious periods of different individuals to be independent and identically distributed. During the infectious period, an infective individual can make contact with another specific individual at a time that is supposed to be from a time-homogeneous Poisson process with intensity β/m. If the latter individual is susceptible, then they will immediately become infected. If we denote with $S(t)$, $I(t)$, and $R(t)$ respectively the number of individuals Susceptible, Infected, and Recovered at time t, then the stochastic process $\{(S(t), I(t), R(t)); t \geq 0\}$ is a Markov process only if the infectious period has the lack-of-memory property. For this reason, despite it is not the most realistic assumption, one particular choice of the time distribution, which guarantees the Markovian property to hold, is the exponential one.

The transition model can be summarized by a CRN reaction table that sums up the possible transition from state $(S(t), I(t), R(t)) = (s, i, r)$ and the respective rates, as shown in Table 2.1.

TABLE 2.1: SIR reactions scheme.

reaction	from	to	rate
$r_1 : s + i \to 2i$	(s, i, r)	$(s-1, i+1, r)$	$\beta si/m$
$r_2 : i \to r$	(s, i, r)	$(s, i-1, r+1)$	μi

If (s, i, r) is the current state of the system, with transition rate $\beta si/m$ the system will evolve in $(s, i+1, r)$ through reaction r_1, and with reaction rate μi reaction r_2 will occur, leading to state $(s, i-1, r+1)$. The probability of r_1 or r_2 occurring, are

$$\mathbb{P}(r_1) = \frac{\beta s}{\beta s + m\mu}, \qquad \mathbb{P}(r_2) = \frac{m\mu}{\beta s + m\mu}. \tag{2.3}$$

Let us consider a single infectious individual and let X be the number of people he will infect before his recovery. Assuming β and s to be approximatively constant over the short time range of a single recovery, X is a geometric distributed variable

$$X \sim Geo\left(p = \frac{m\mu}{\beta s + m\mu}\right), \tag{2.4}$$

where we consider a new infection as a failure and the recovery as success, so the probability of success of the Geometric distribution coincides with $\mathbb{P}(r_2)$. We can directly derive the mean number of people infected by a single infected individual as the mean of the geometric distribution $\mathbf{E}[X] = \beta S/m\mu$, which gives the reproductive number for the fixed time with status (s, i, r). It should be noted that, if $S \simeq m$, then $\rho = \beta/\mu$, which is the well-known basic reproduction number (indicated with ρ^0) associated with the initial status of the pandemic, namely the number of individuals that a single infectious person can infect before they recover. All the stochastic theory behind the SIR model can be translated into the deterministic SIR model when the population size m is large enough [AK15]: using the property of Markov processes, we can derive the deterministic and diffusion approximations for the whole trajectories $S(t), I(t), R(t)$ describing the number of people in each of the compartments, for each time t. The deterministic version of the SIR model is determined by the following Ordinary Differential Equations (ODE) system :

$$\begin{cases} \dfrac{dS(t)}{dt} = -\dfrac{\beta S(t) I(t)}{m}, \\ \dfrac{dI(t)}{dt} = \dfrac{\beta S(t) I(t)}{m} - \mu I(t), \\ \dfrac{dR(t)}{dt} = -\mu I(t). \end{cases} \tag{2.5}$$

From the stochastic model, we can also derive the interpretation of the parameters: in particular, $1/\mu$ gives the mean number of days of the transition from I to R.

2.2 SIPRO model

The classical SIR model divides the population into three different compartments: Susceptibles, Infectives, and Recovered (including deads). However, for a virus as the Sars-Cov2, it seems inappropriate to compare official data on infected and recovered people with the curves predicted by the SIR model at least for two reasons [Mul21]:

1. the infection is to a large extent carried by asymptomatic people that are normally not counted in official records;

2. people that tested positive, and hence counted as infected, are usually quarantined and should not be considered as a source of contagion.

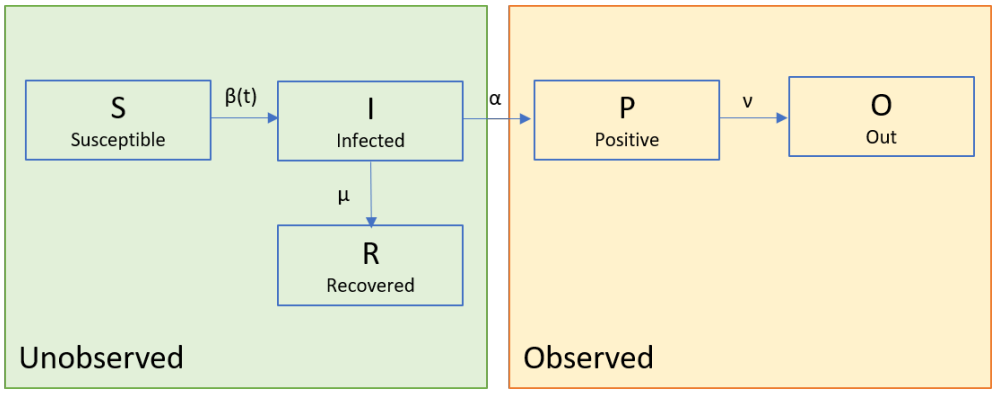

FIGURE 2.2: Graphical representation of SIPRO model.

We propose a new model that can account for these two facts, but that, at the same time, remains sufficiently tractable to be fit using public data. In our model, as in the SIR, after having been in contact with an infected person, Susceptible individuals immediately become Infected themselves. However, only some of the infected are diagnosed by a test and then enter the Positive compartment. In this case, they stop spreading the virus, since we assume they are isolated. People that have been counted as positive finally enter the Out compartment once they either recover or die. Those infected that remain undetected, will eventually recover or die, entering the Recovered compartment. The population is then divided into five different compartments (see Figure 2.2):

- Susceptibles (S);
- Infectives (I);
- Positives (P), individuals that have been infected, and after testing positive, are quarantined or isolated;
- Recovered (R), individuals that have been infected, and recover or die without having been tested;
- Out (O), recovered or dead people who were accounted as positive.

The ODE system associated with SIPRO model is the following

$$\begin{cases} \dfrac{dS(t)}{dt} = -\dfrac{\beta(t)}{m} I(t)S(t), \\ \dfrac{dI(t)}{dt} = \dfrac{\beta(t)}{m} I(t)S(t) - \mu\, I(t) - \alpha\, I(t), \\ \dfrac{dP(t)}{dt} = \alpha\, I(t) - \nu\, P(t), \\ \dfrac{dR(t)}{dt} = \mu\, I(t), \\ \dfrac{dO(t)}{dt} = \nu\, P(t). \end{cases} \tag{2.6}$$

We name this model SIPRO-$(\beta(t), \alpha, \mu, \nu)$. As for the SIR model, our dynamical system is the large population limit of a stochastic model; the chemical reaction network which describes the stochastic model is composed of five species $\{S, I, P, R, O\}$, six

complexes $\{S+I, I, 2I, P, R, O\}$, and four reactions $\{r_1, r_2, r_3, r_4\}$ in the reaction table below (see Table 2.2). Moreover, the stochastic representation of the model allows us to interpret the parameters: $1/\nu$ can be interpreted as the mean time an individual spends in the positive compartment, while $1/\alpha$ is the mean time between the infection and its detection, and $1/\mu$ is the mean time it takes to recover for a person whose infection has never been detected. These interpretations can be carried over to the deterministic limit model.

The function $\beta(t)$ denotes the rate of infection as a function of time. It varies over time due to changes in population behavior in response to containment measures like social distancing or mask-wearing, as well as the spread of new viral strains with different characteristics.

To derive the reproductive number for the SIPRO model, we first introduce the transition scheme for the stochastic model in Table 2.2.

Let (s, i, p, r, o) be the current state of the system at time t. The system can evolve in $(s-1, i+1, p, r, o)$, $(s, i-1, p+1, r, o)$, $(s, i-1, p, r+1, o)$, $(s, i, p-1, r, o+1)$ for reactions r_1, r_2, r_3 or r_4, respectively, with reactions rate that are $\beta(t)si/m$, αi, μi and νp. An infectious individual can lead to a new infection, through reaction r_1, or can become positive/recovered through reaction r_2 or r_3, with

$$\mathbb{P}(r_1) = \frac{\beta(t)s}{\beta(t)s + (\mu+\alpha)m} \quad \text{and} \quad \mathbb{P}(r_2 \cup r_3) = \frac{(\mu+\alpha)m}{\beta(t)s + (\mu+\alpha)m,}. \tag{2.7}$$

Let us consider a single infectious individual and let X be the number of people he will infect before his recovery. Assuming β and s to be approximatively constant over the short time range of a single recovery, X is a geometric distributed variable

$$X \sim Geo\left(p = \frac{(\mu+\alpha)m}{\beta(t)s + (\mu+\alpha)m}\right). \tag{2.8}$$

We can directly derive the mean number of people infected by a single infected individual as

$$\mathbf{E}[X] = \frac{1-p}{p} = \frac{\beta(t)s}{(\mu+\alpha)m}. \tag{2.9}$$

Observe that, for $t = 0$, the number s of susceptible individual is almost equal to m, so that $\rho^0 = \beta(0)/(\mu+\alpha)$, which is the basic reproduction number. Evaluating the expected number of secondary cases $\mathbb{E}(X)$ for each time t characterized by $S(t)$ susceptibles, we can define the *effective reproduction number* for the SIPRO model as

$$\rho^{\text{eff}}(t) = \frac{\beta(t)S(t)}{(\mu+\alpha)m}. \tag{2.10}$$

Despite the SIPRO carries more information with respect to the original SIR model,

TABLE 2.2: SIPRO reactions scheme.

reaction	from	to	rate
$r_1 : s+i \to 2i$	(s, i, p, r, o)	$(s-1, i+1, p, r, o)$	$\beta(t)si/m$
$r_2 : i \to p$	(s, i, p, r, o)	$(s, i-1, p+1, r, o)$	αi
$r_3 : i \to r$	(s, i, p, r, o)	$(s, i-1, p, r+1, o)$	μi
$r_4 : p \to o$	(s, i, p, r, o)	$(s, i, p-1, r, o+1)$	νp

it remains dramatically simplified: it does not account for many features, as the incubation period, the weekly pattern of the new infected caused by the decreased testing on weekends, the possibility of reinfections, the age structure of the population, the non-uniformity of the contact network. Its main advantage is that it remains a tractable model.

2.3 Analyzing the COVID-19 pandemic in Italy with SIR and SIPRO models

In this section, we present the statistical models in which we combine SIR and SIPRO structures with the available data, to construct the algorithm used to make inference. Using both models, we use the daily proportions of positive tested people and recovered recorded individuals in each region i, derived from the Protezione Civile database [DPCa], dividing the daily counts by the total population of each region (that we consider constant across the analysis). Note that the theory explained in Sub-section 2.1.2 and Section 2.2 still holds if we substitute counts with proportions. Let T denote the number of days analyzed, while $n = 21$ is the number of regions considered. We first present the performances of SIPRO and SIR mixed models on simulations to better understand the practical applicability of the proposed model: data were respectively simulated from SIPRO and SIR models, and estimated with the relative schemes. We then present the real data analysis of the first wave of the pandemic (February 2020 – May 2020).

To the extent of clarifying the estimation of the underlying parameter of the COVID-19 pandemic in Italy, we briefly outline the data collection process employed by the Italian system (see Figure 2.3[2]). The daily data collection process in Italy involves four key steps: (i) within each region, data are recorded on an institutional application by hour 16:30; (ii) the Ministry of Health verifies the data and transmits them to the Department of Civil Protection (DPC) by hour 17:30; (iii) the DPC thoroughly analyzes and processes the dataset to standardize it by hour 18:00; finally, (iv) the data are uploaded onto GitHub and the Dashboard. It's worth noting that the execution of this plan was influenced by the evolution of the pandemic, resulting in variations during different phases. However, focusing on data collected during the first wave (February 2020-July 2020), we can reasonably consider the regional internal policy to be sufficiently homogeneous.

2.3.1 SIPRO statistical model to analyze Italian data: a simulation study

We assume that the epidemic evolution in the i-th region is described by an independent SIPRO-$(\beta_i(t), \alpha_i, \mu, \nu_i)$ model, neglecting contacts between the populations in different regions. However, the SIPRO compartments cannot be directly observed. Instead, the available data comprises a noisy version of the proportions of Positive and Out individuals, denoted as $y_{P,i}(t)$ and $y_{O,i}(t)$, respectively, which we envision as a realization of the random variables $Y_{P,i}(t)$ and $Y_{O,i}(t)$. We can introduce a third complementary component $Y_{C,i}(t) = 1 - Y_{P,i}(t) - Y_{O,i}(t)$ that makes our state vector $\mathbf{Y}_i(t) = (Y_{P,i}(t), Y_{O,i}(t), Y_{C,i}(t))$ naturally confined in the simplex $\{y \in [0,1]^3 : \|y\|_1 = 1\}$.

Our choice is to model $\mathbf{Y}_i(t)$ using a Dirichlet distribution centered at $(p_i(t), o_i(t), s_i(t) + \imath_i(t) + r_i(t))$ with a further parameter γ that rules the noise

[2]License creative commons attribution 4.0 international

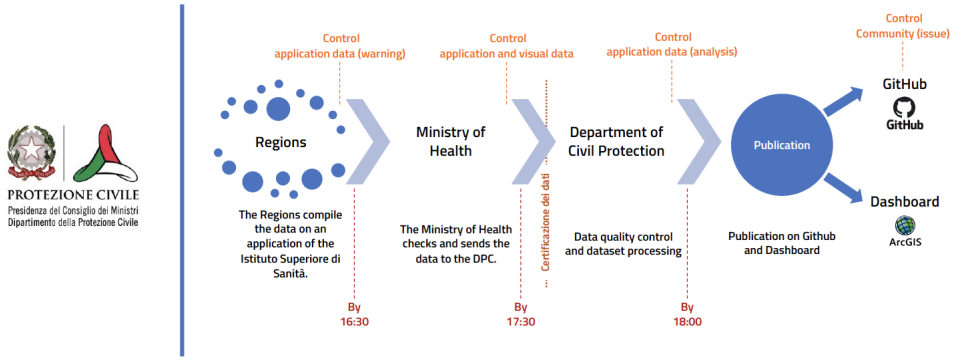

FIGURE 2.3: Italian mechanism for the collection and storage of the Covid daily data (see [DPCb]).

amplitude and that we consider as common to all regions. The precise formulation is given in Equation (2.14).

The assumption of no communication between regions is overall unrealistic, except for the period of the national lock-down (from March 9 to May 4, 2020), but public data do not contain any piece of information about the traffic flows, and it would be impossible to estimate the contribution of people getting sick in a region and infecting other people in a different one. While we know that Sars-Cov-2 can reinfect its host due to the waning of immunity, this effect can be neglected, since the time period considered in this study is short enough. The Italian regions have a significant heterogeneity: their communication networks, containment measures, testing policies, health systems, and availability of ICU beds are different and the quantities $(\beta_i(t), \alpha_i, \nu_i)$ are affected by such heterogeneity. We consider them as individual characteristics of each region. In particular, to help gather common information, we assume that the ν_i are drawn for a common lognormal population (with parameters ν and ω_ν, as in Equation (2.16)) while, after some preliminary test, we prefer not to impose any common distribution on the α_i. The parameter μ is related to the natural duration of the contagion of the undetected cases, we assume it to be the same all over the countries. To simplify inference, the functions $\beta_i(t)$ are selected into the parametric family of continuous linear interpolations between the values $\beta_{ik} = \beta_i(t_k)$ at pre-specified equi-spaced days $\{t_k\}_{k=1}^K = (k-1)\Delta$, where Δ is the number of days between each node of the time grid and the following one. Moreover, we impose that $\beta_i(t)$ is constant in the last time interval of the grid, namely between day t_{K-1} and day $t_K = T - 1$. Indeed, identifying a change in the reproduction number in the latest period is practically impossible since its effect would be visible only later.

The values at the nodes, $\{\beta_{ik}\}_{\substack{i=1\dots n\\ k=1\dots K}}$, are regional parameters to be estimated, and we assume they are drawn from a common lognormal population with parameters $(\beta(k), \omega_\rho)$ to be estimated. The solutions of each regional ODE system in Equation (2.6), are uniquely determined by the parameters that we have introduced, and the initial conditions $x_i(0) = (s_i(0), \imath_i(0), p_i(0), r_i(0), o_i(0))$ that have to be estimated as well. Let us denote by

$$\boldsymbol{\phi} = \left(\{x_i(0)\}_{i=1,\dots,n}, \{\alpha_i\}_{i=1,\dots,n}, \{\nu_i\}_{i=1,\dots,n}, \{\beta_{ik}\}_{\substack{i=1,\dots,n\\ k=1,\dots,K}} \right) \tag{2.11}$$

the vector of the *individual* (regional) parameters and by

$$\boldsymbol{\theta} = (\log \nu, \omega_\nu, \mu, \gamma, \{\log \beta(k)\}_{k=1,\dots,K}, \omega_\beta) \tag{2.12}$$

the vector of the *global* parameters. Adopting a Bayesian approach, our statistical model is defined through the joint probability density

$$f(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{t=0}^{T} f(y_{P,i}(t), y_{O,i}(t), y_{C,i}(t) | \phi_i, \theta) f(\phi_i | \theta) f(\theta), \tag{2.13}$$

where $\boldsymbol{y} = \{y_{P,i}(t), y_{O,i}(t), y_{C,i}(t)\}_{\substack{i=1\dots n \\ t=0\dots T}}$, with $y_{C,i}(j) = 1 - y_{P,i}(j) - y_{O,i}(j)$, for all i and t, $\phi_i = (x_i(0), \alpha_i, \nu_i, \{\beta_{ik}\}_{k=1,\dots,K})$, and the conditional probability densities in the right-hand side are specified as follows, for $i = 1, \dots, n$,

$$Y_{P,i}(t), Y_{O,i}(t), Y_{C,i}(t) | \phi_i, \theta \overset{i.i.d.}{\sim} Dirichlet\big(\gamma\, p_i(t),\, \gamma\, o_i(t),\, \gamma\,(1 - p_i(t) - o_i(t))\big), \tag{2.14}$$

$$\log(\beta_{ik}) | \theta \overset{i.i.d}{\sim} \mathcal{N}(\log(\beta_k), \omega_\beta^2), \quad k = 1, \dots, K, \tag{2.15}$$

$$\log(\nu_i) | \theta \overset{i.i.d}{\sim} \mathcal{N}(\log(\nu), \omega_\nu^2), \tag{2.16}$$

$$\frac{1}{\alpha_i} \overset{i.i.d}{\sim} \mathcal{U}(1, 30), \tag{2.17}$$

while

$$\big(s_i(0) + r_i(0), \imath_i(0), p_i(0), o_i(0)\big) | r_i(0) \overset{i.i.d}{\sim} Dirichlet(4, 2, 2, 2), \tag{2.18}$$

$$r_i(0) \overset{i.i.d.}{\sim} \mathcal{U}(0, 0.001), \tag{2.19}$$

and $f(\theta)$ are the appropriate prior distributions we now define.

To infer the parameters of our model we use a Bayesian approach, with the following prior distributions

$$f(\theta) = f(\nu) f(\omega_\nu) f(\mu) f(\gamma) f(\omega_\beta) \prod_{k=1}^{K} f(\beta(k)), \tag{2.20}$$

where the marginal priors of each parameter are

$$\log \beta(k) \sim \log \gamma \sim \log \nu \sim \log \mu \sim \mathcal{N}(0, 1000), \tag{2.21}$$

$$\omega_\beta^2 \sim \omega_\nu^2 \sim \mathcal{IG}(0.1, 0.1). \tag{2.22}$$

Priors (2.18), (2.21), and (2.22) carry very little information; in particular, (2.21) and (2.22) are standard weakly-informative choices commonly used for mean and variance of random effects. On the contrary, on the other parameters, some prior information is encapsulated in the prior densities. In particular, we assume the transition Infected-Positive time (namely $1/\alpha_i$) is between 1 to 30 days (prior (2.17)). In addition, we treat $r_i(0)$ in a slightly different way compared with other initial conditions: as we expect it to be hardly identifiable, we constrain it in a reasonable and realistic interval, assuming that at the beginning of the pandemic, the percentage of recovered individuals was smaller than 0.1%. As can be observed from Equation (2.19), it remains in a small interval around zero so that it does not strongly impact the dynamics.

Our proposed model was tested on simulations, which helped us understand its

strengths and weaknesses. In our first contribution [ABM21], we used a Metropolis-within-Gibbs algorithm, obtaining quite good results but facing some identifiability problems, despite the simplicity of the proposed model. Thus, we re-approach here the estimation by combining the implemented Metropolis-within-Gibbs algorithm with parallel tempering (see Chapter 1). The main problem concerns the recovery transition from Infected to Recovered compartment: estimation of $r_i(0)$ and μ is very challenging and negatively reflects also on the estimation of other parameters of the model. This can be intuitively justified by the fact that the cumulative number of Recovered individuals at day t is not really impacting the evolution of the pandemic (and thus the estimation of the Observed part of the pandemic which enters directly into the likelihood): small fluctuations of the individual parameters α_i and β_{ik} around their true values can account for (small) deviations of μ from the true values. The higher the information of the dataset under analysis, the lower the error on the final parameters estimates.

The code, implemented in `Julia 1.8.5`, is composed of 6000 iterations of PT (out of which 5000 discarded in burning), each composed of 1000 MCMC repetitions with 990 iterations discarded by burn-in, and the last one used to perform PT swap. For each region i, at each iteration b, all the individual parameters are proposed using a Metropolis-Hastings update. In particular, on $(\alpha_i, x_i(0))$ we use a multivariate adaptive random walk proposal, on ν_i an adaptive univariate random walk proposal, and, finally, on the nodes $\{\beta_{ik}\}_{\substack{i=1,\dots,n \\ k=1,\dots,K}}$ we use again a multivariate adaptive random walk proposal. All proposals' standard deviations are adapted using method 4 in [AT08], with hyperparameters $\gamma^{b+1} = 1/(b+1)^{0.7}$ and $\alpha^* = 0.25$. Each time we propose a new parameter we solve the ODE system numerically with the Euler discretization scheme with a unit time step and compute the likelihood. All global parameters are updated with standard Gibbs or Metropolis-Hastings updates.

We present here the result obtained on a simulated dataset. To better show the low-identifiability problems, we present the results obtained on the same simulated dataset (1) estimating all the parameters and (2) keeping μ fixed to the real value (e.g., the one used to simulate the data). We report the percentage of individual parameters correctly estimated (true values is in the 95% credible interval) in Table 2.3 and the posterior 95% credible interval with the true value of global parameters in Table 2.4. Moreover, we report in Figure 2.4 the chain of the parameter μ, in the case in which we try to estimate it, to show that the algorithm almost reaches the true value (red solid line) but does not correctly identify it. For lack of space, we avoid

TABLE 2.3: Simulated SIPRO dataset - Results on individual parameters.

Parameter	Percentage (μ estimated)	Percentage (μ fixed)
$\iota_i(0)$	80.95	95.24
$p_i(0)$	47.62	47.62
$r_i(0)$	0.00	0.00
$o_i(0)$	28.57	23.81
α_i	76.19	95.24
ν_i	90.48	90.48
$\beta_i(t)$	92.55	76.30
$\rho_i^{\text{eff}}(t)$	82.35	95.13

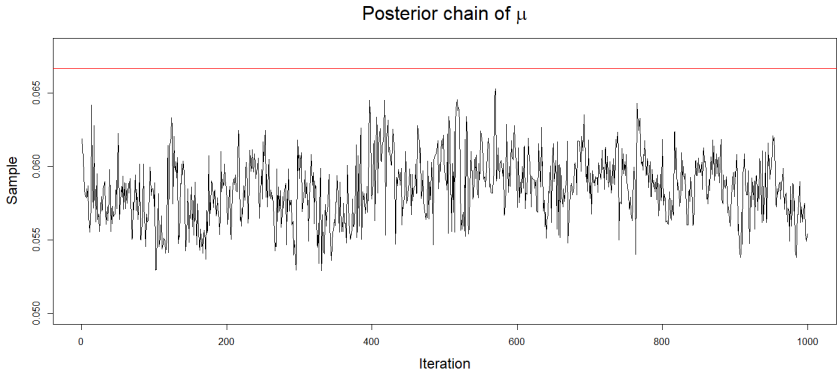

FIGURE 2.4: Chain of parameter μ (black). The horizontal red line indicates the true value of the parameter.

reporting the fitted trajectories for each region and each compartment, but we define an adapted-MSE time-dependent index to give an overall idea of the goodness of fit of the SIPRO model.

Definition 2.3.1. *Given the B posterior sampled trajectories* $\{\hat{X}_i^b(t)\}_{b=1}^B$*, for each patient i, and the true trajectories* $\{X_i(t)\}$*, the overall* $MSE_{adp}^X(t)$ *time dependent index is*

$$MSE_{adp}^X(t) = \sum_{i=1}^{n} \sum_{b=1}^{B} \frac{(\hat{X}_i^b(t) - X_i(t))^2}{nB}. \tag{2.23}$$

The adapted-MSE plot, for each compartment, is shown in Figure 2.5: the shape of the trajectories and of the index is quite similar, as higher errors correspond to higher proportions of people in the compartment. We also evaluate, for each trajectory of interest, the percentage of times t such that the true value is in the posterior 95% credible interval (Table 2.5).

An overall view of the reported results shows that fixing the parameter μ largely improves the estimation (in particular, the Infected part of the SIPRO model): the percentage of correctly estimated individual α_i goes from 76.19% to 100%, $\imath_i(0)$ from 85.71% to 90.48%. A huge improvement can also be seen in the trajectory fit: the percentage of correctly identified time points increase in all trajectories (for $\imath(t)$ from 64.35% to 92.86%). Parameter $r_i(0)$ still remains practically non-identifiable, but this

TABLE 2.4: Simulated SIPRO dataset - Results on global parameters.

	μ estimated				μ fixed		
Parameter	2.5%	Mean	97.5%	Real	2.5%	Mean	97.5%
$\log(\nu)$	-3.203	-2.997	-2.795	-3.045	-3.211	-3.000	-2.796
ω_ν	0.334	0.453	0.645	0.500	0.330	0.451	0.622
γ	5.859	5.915	5.967	6.000	5.846	5.902	5.957
μ	0.055	0.059	0.063	0.067	-	-	-
ω_β	0.091	0.156	0.735	0.100	0.084	0.113	0.153
$\log(\beta_1)$	-0.772	-0.154	-0.019	0.000	-0.638	-0.089	0.015
$\log(\beta_2)$	-1.061	-0.899	-0.816	-0.916	-1.048	-0.893	-0.827
$\log(\beta_3)$	-1.064	-0.911	-0.640	-0.916	-0.928	-0.837	-0.692
$\log(\beta_4) = \log(\beta_5)$	-1.921	-1.329	-0.449	-1.204	-1.119	-0.800	0.112

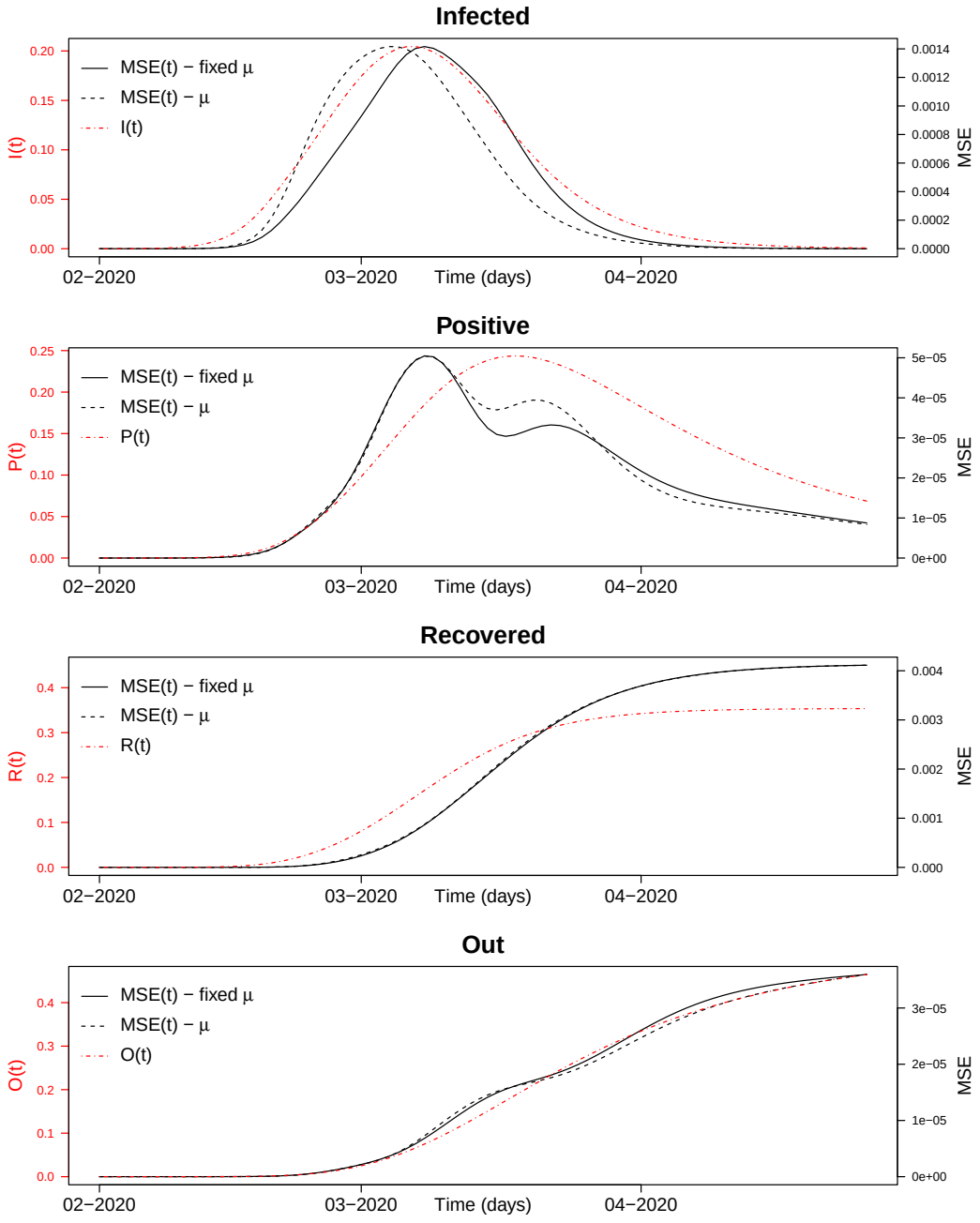

FIGURE 2.5: Mean time-dependent MSE index (black-solid line and black-dotted line) to evaluate the fit of SIPRO-mixed-model, for Infected, Positive, Recovered, and Out trajectories, respectively with fixed $\mu = 1/5$ and μ estimated (see MSE Definition 2.3.1). The red line is the number of Infected, Positive, Recovered, and Out in Italy for each day. The y-axes values of the red line are shown on the left-side, while the ones of the black lines are on right-side.

does not impact the overall performance of the method, as it only affects the Removed part of the epidemic. Moreover, we note that the posterior density for $r_i(0)$ coincides with the uniform prior we use. As the main goal of the SIPRO model is to describe the asymptomatic part of the population that can transmit the infection without being part of the daily counts, and, thus, estimate the effective reproduction number taking into account that most infections are carried on by them, the results obtained give positive feedback on the model applicability. However, some warnings and remarks need to be pointed out: the model identifiability (even when we keep μ fix) is quite unstable, and particular parameter configurations can lead to a total non-identifiability of the model. This, in particular, may happen when $s_i(t)$ or $\iota_i(t)$ are almost zero, leading to a vanishing contribution of the transition parameter $\beta_i(t)$ that becomes non-identifiable. For this reason, long periods of controlled pandemic, with almost no cases, should not be part of the analysis, as results are not reliable anymore. Moreover, as the impact of huge changes in the effective reproduction number is delayed in time (which can be clearly seen by the time of effectiveness of social distances measures), the last part of functions $\beta_i(t)$ is more difficult to estimate.

Full Simulation Study

To assess the applicability of the SIPRO model and its corresponding MCMC code, we conduct numerical tests across five additional scenarios. In what follows, we refer to the case study presented in Section 2.3.1 as scenario S1. These scenarios, distinguished by varying parameter values, are visually depicted in Figures 2.6 to 2.9. Changes include adjusting measurement noise, altering the level and/or timing of the epidemic peak, and modifying the initial conditions levels. The performance summaries of the algorithm are presented in Tables 2.6 to 2.7. As for the intrinsic structure of the SIPRO model, parameters $p_i(t)$, $o_i(t)$, and ν_i, whose estimation directly relies on the collected data, are well-estimated across all analyzed scenarios. However, low-identifiability issues affect other parameters. Initial conditions ($\iota_i(0)$, $p_i(0)$, $o_i(0)$), typically low, pose challenges in estimation. Scenario S5 provides insight into the impact of higher initial condition values, indicating that higher initial conditions lead to improved estimation. Furthermore, we employ scenarios S1-S6 to test the identifiability of $\beta_i(t)$ (and $\rho_i^{\text{eff}}(t)$) in different epidemic phases. Scenarios S1-S4 represent ongoing epidemics with a sufficient infected proportion for estimating $\beta_i(t)$. Scenario S5 differs a little bit; it presents a situation when right after a high peak of infections, the number of infected slowly decreases and almost reaches zero: the number of infected is still high enough to give the necessary information to estimate $\beta_i(t)$. Finally, Scenario S6, with infections rapidly decreasing to zero, indicates a loss of identifiability for $\beta_i(t)$. Additionally, S1-S6 are utilized to examine the

TABLE 2.5: Trajectories percentage of correctly estimated values over all times t.

Parameter	Percentage (μ estimated)	Percentage (μ fixed)
$\iota_i(t)$	71.21	95.90
$p_i(t)$	83.94	87.38
$r_i(t)$	71.87	81.45
$o_i(t)$	82.23	84.22
$s_i(t)$	76.36	86.60

impact of measurement noise (e.g., parameter γ) on estimation quality. As expected, lower measurement noise leads to more accurate parameter and trajectory estimations. In line with the earlier discussion in Section 2.3.1, simulations underscore the persisting non-identifiability of $r_i(0)$. The posterior density is still identical to the prior, implying that, when the simulated value falls outside the prior 95% interval, it will not count in the percentage of the estimated parameter. Nevertheless, the simulation study provides an overview of potential scenarios, offering guidelines on real-world occurrences, how to address such situations, and which parameters of interest yield reliable results. It specifically emphasizes that estimating the effective reproduction number and the proportion of infective individuals becomes challenging during low-incidence transition periods, like the interval between one peak and the following (e.g., summer 2020 in Italy). Therefore, additional attention should be given when modeling these situations. From a practical standpoint, challenges arise when estimating the random effect parameters ($\log \beta_k$) for time-nodes t_k in this low-incidence time window. A possible solution is to set only two nodes for the entire period, defining the beginning and the end, thus reducing the number of parameters reliant on a period with insufficient information. To further investigate the role of μ, we test each scenario (S1-S6) with and without estimating it: fixing μ to its simulation value yields overall better performance, as previously highlighted. Given that the simulation study enforces the importance of carefully handling μ in the SIPRO model, Section 2.3.1 provides additional insights into the model identifiability of μ.

TABLE 2.6: Percantage of correctly estimated individual parameters, under different scenarios S1-S6, with μ as a parameter to be estimated (μ) or μ fixed to the true value used to simulate the corresponding dataset.

Param.	S1 (μ)	S1	S2 (μ)	S2	S3 (μ)	S3	S4 (μ)	S4	S5 (μ)	S5	S6 (μ)	S6
$\iota_i(0)$	80.95	95.24	85.71	90.48	57.14	76.19	90.48	90.48	80.95	80.95	95.24	100
$p_i(0)$	47.62	47.62	95.24	95.24	66.67	61.9	9.52	9.52	95.24	95.24	95.24	95.24
$r_i(0)$	0	0	0	0	0	0	0	0	0	0	0	0
$o_i(0)$	28.57	23.81	76.19	76.19	23.81	23.81	33.33	33.33	90.48	90.48	66.67	66.67
α_i	76.19	95.24	66.67	85.71	61.90	100	85.71	100	90.48	61.90	85.71	95.24
ν_i	90.48	90.48	100	100	95.24	95.24	95.24	95.24	90.48	95.24	95.24	100
$\iota_i(t)$	71.21	95.20	70.10	88.43	60.74	92.30	77.13	94.41	89.31	82.00	42.97	49.56
$p_i(t)$	83.94	87.38	94.63	97.07	88.37	89.99	89.94	91.86	94.08	94.30	71.54	73.53
$r_i(t)$	71.87	81.45	74.20	80.12	77.91	84.33	84.94	89.42	84.61	84.27	83.00	82.83
$o_i(t)$	82.23	84.22	90.75	91.25	81.67	84.05	84.33	85.05	97.67	97.51	91.09	91.14
$s_i(t)$	76.36	86.60	84.16	82.22	76.25	88.87	86.32	93.96	90.53	89.81	82.78	82.28
$\rho_i^{\text{eff}}(t)$	82.35	95.30	48.85	94.29	89.92	95.24	95.57	95.69	93.67	86.78	44.37	45.55
$\beta_i(t)$	92.55	76.30	13.00	81.57	89.58	57.48	96.30	88.18	89.52	96.36	45.55	50.25

TABLE 2.7: Quantile of global parameters, in different scenario S1-S6, with μ as a parameter to be estimated (μ) or μ fixed to the true value.

	μ estimated				μ fixed		
Parameter S2	2.5%	Mean	97.5%	Real	2.5%	Mean	97.5%
$\log(\nu)$	-1.672	-1.422	-1.178	-1.609	-1.678	-1.418	-1.151
ω_ν	0.432	0.587	0.793	0.500	0.441	0.594	0.826
γ	9.919	9.970	10.013	10.00	9.911	9.961	10.009
μ	0.227	0.264	0.303	0.200	-	-	-
ω_β	0.172	0.285	0.551	0.100	0.116	0.167	0.304
$\log(\beta_1)$	-0.919	-0.208	-0.007	0.000	-0.272	-0.067	0.028
$\log(\beta_2)$	-1.597	-1.282	-0.524	-0.916	-1.163	-1.031	-0.924
$\log(\beta_3)$	-1.869	-1.382	-0.416	-0.916	-1.102	-0.987	-0.880
$\log(\beta_4) = \log(\beta_5)$	-3.089	-1.990	-0.090	-1.204	-1.522	-1.356	-1.225
Parameter S3	2.5%	Mean	97.5%	Real	2.5%	Mean	97.5%
$\log(\nu)$	-3.100	-2.842	-2.594	-3.045	-3.097	-2.858	-2.604
ω_ν	0.424	0.579	0.810	0.50	0.428	0.574	0.803
γ	4.831	4.887	4.939	5.00	4.818	4.880	4.936
μ	0.054	0.058	0.058	0.067	-	-	-
ω_β	0.107	0.206	0.629	0.10	0.096	0.157	0.698
$\log(\beta_1)$	-0.854	-0.215	-0.037	0.00	-0.404	-0.092	0.005
$\log(\beta_2)$	-1.159	-0.852	-0.727	-0.916	-0.995	-0.842	-0.760
$\log(\beta_3)$	-1.892	-1.094	-0.802	-0.916	-1.030	-0.901	-0.741
$\log(\beta_4) = \log(\beta_5)$	-1.623	-1.020	-0.612	-1.204	-1.021	-0.739	-0.304
Parameter S4	2.5%	Mean	97.5%	Real	2.5%	Mean	97.5%
$\log(\nu)$	-3.090	-2.853	-2.611	-3.045	-3.115	-2.859	-2.592
ω_ν	0.424	0.581	0.789	0.50	0.427	0.591	0.834
γ	5.866	5.917	5.968	6.00	5.852	5.908	5.959
μ	0.061	0.064	0.066	0.067	-	-	-
ω_β	0.103	0.187	0.587	0.10	0.098	0.151	0.484
$\log(\beta_1)$	-1.187	-0.471	-0.293	-0.357	-0.562	-0.398	-0.285
$\log(\beta_2)$	-1.485	-0.673	-0.483	-0.693	-0.727	-0.627	-0.484
$\log(\beta_3)$	-1.406	-0.919	-0.718	-0.916	-1.071	-0.84	-0.623
$\log(\beta_4) = \log(\beta_5)$	-2.996	-0.946	-0.382	-0.916	-0.931	-0.580	0.095
Parameter S5	2.5%	Mean	97.5%	Real	2.5%	Mean	97.5%
$\log(\nu)$	-1.668	-1.420	-1.168	-1.609	-3.115	-2.859	-2.592
ω_ν	0.432	0.581	0.795	0.500	0.427	0.591	0.834
γ	9.963	10.010	10.060	10.000	5.852	5.908	5.959
μ	0.047	0.048	0.048	0.048	-	-	-
ω_β	0.100	0.202	1.338	0.100	0.098	0.151	0.484
$\log(\beta_1)$	-1.641	-0.790	-0.623	-0.693	-0.562	-0.398	-0.285
$\log(\beta_2)$	-2.849	-1.013	-0.616	-0.916	-0.727	-0.627	-0.484
$\log(\beta_3)$	-5.011	-1.536	-0.658	-0.916	-1.071	-0.84	-0.623
$\log(\beta_4) = \log(\beta_5)$	-1.399	-0.901	0.352	-1.204	-0.931	-0.580	0.095
Parameter S6	2.5%	Mean	97.5%	Real	2.5%	Mean	97.5%
$\log(\nu)$	-1.676	-1.417	-1.152	-1.609	-1.682	-1.423	-1.180
ω_ν	0.433	0.586	0.819	0.500	0.429	0.588	0.777
γ	9.864	9.916	9.965	10.000	9.960	10.009	10.054
μ	0.199	0.200	0.202	0.200	-	-	-
ω_β	0.371	0.530	1.792	0.100	0.106	0.168	0.454
$\log(\beta_1)$	-2.819	-0.092	0.230	0.000	-0.782	-0.697	-0.552
$\log(\beta_2)$	-0.032	0.340	0.581	0.336	-2.123	-0.808	-0.461
$\log(\beta_3)$	-14.000	-7.148	-2.867	0.336	-1.241	-0.970	-0.142
$\log(\beta_4) = \log(\beta_5)$	3.065	3.766	4.465	0.000	-1.175	-0.722	0.742

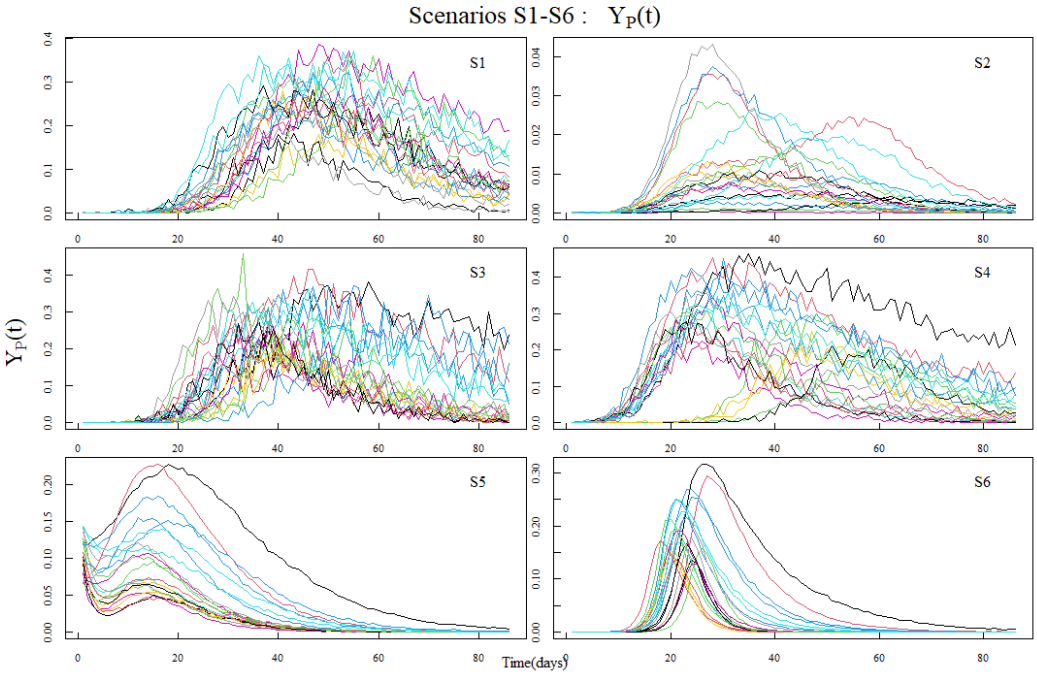

FIGURE 2.6: Regional trajectories $Y_P(t)$ for scenarios S1-S6.

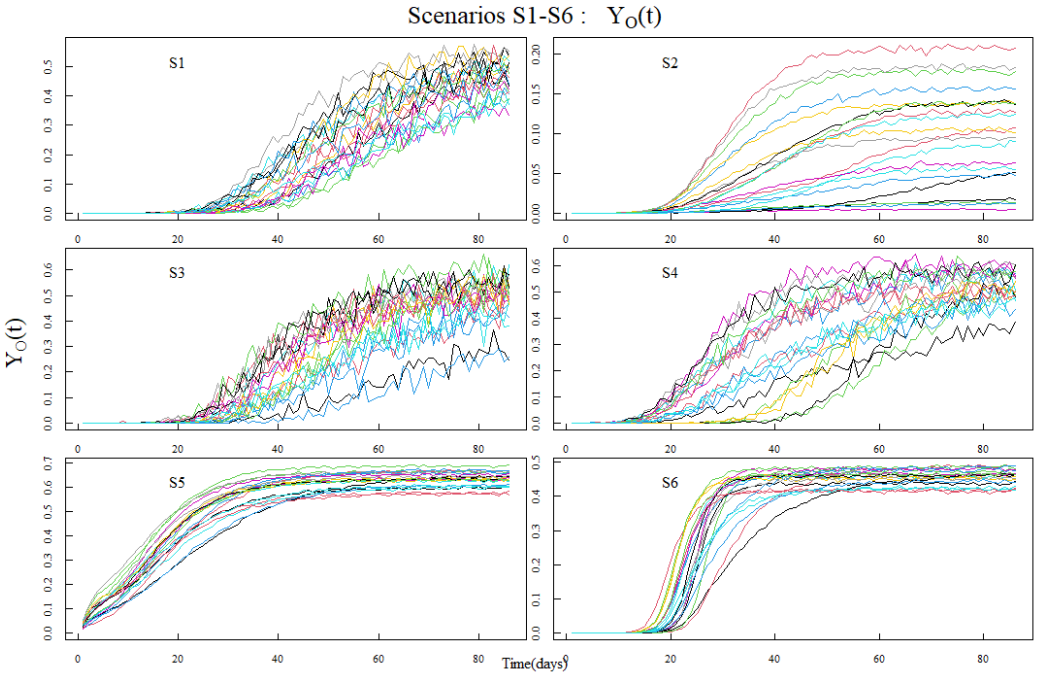

FIGURE 2.7: Regional trajectories $Y_O(t)$ for scenarios S1-S6.

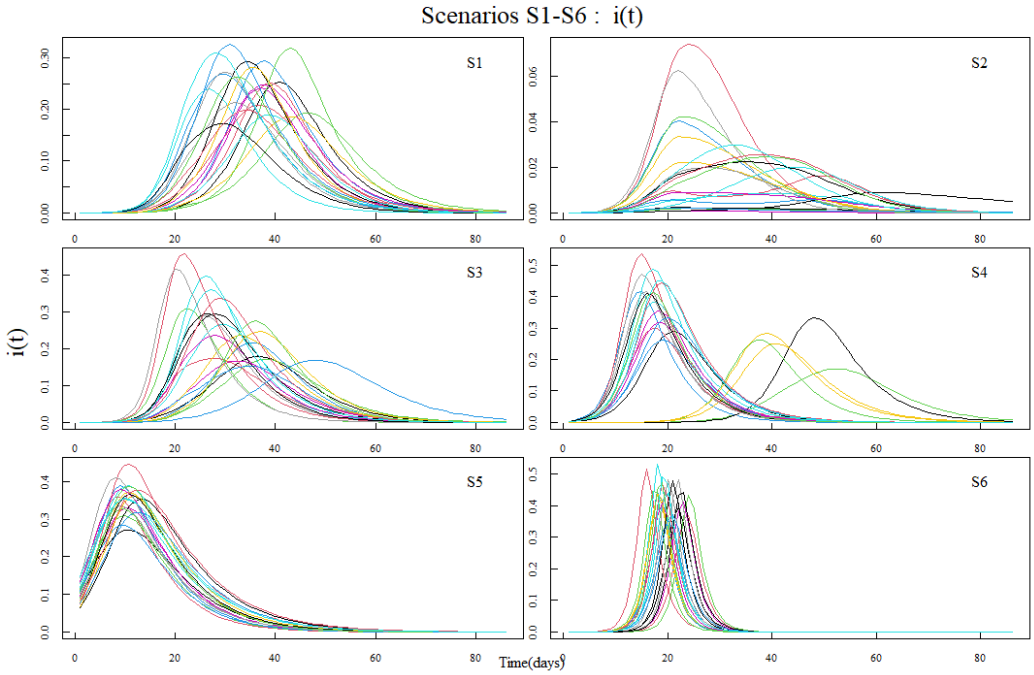

FIGURE 2.8: Regional trajectories $i(t)$ for scenarios S1-S6.

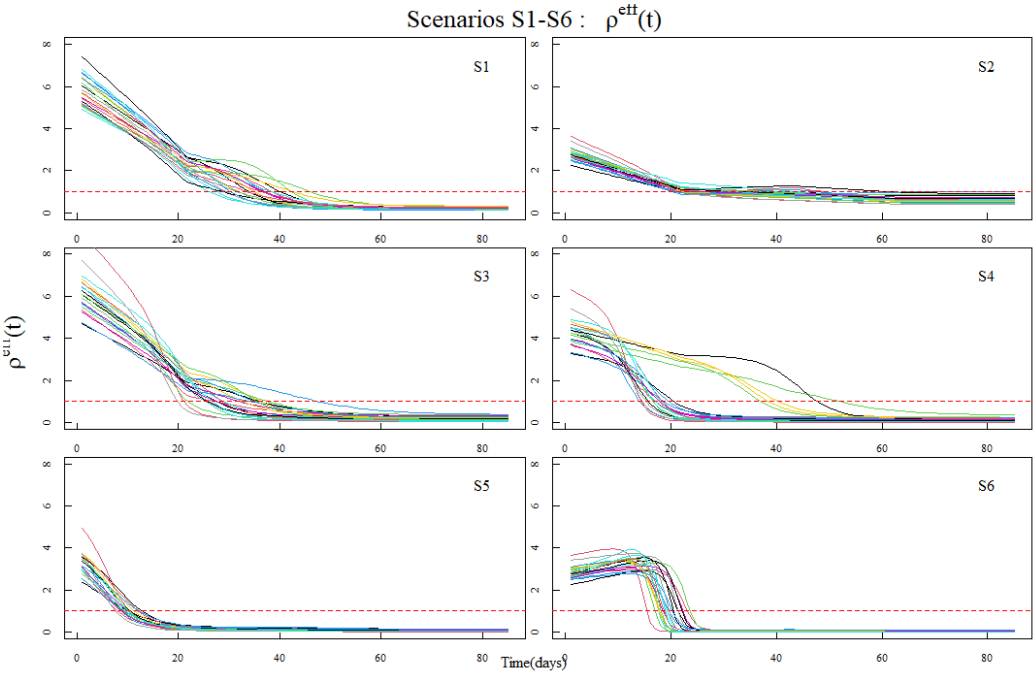

FIGURE 2.9: Regional trajectories $\rho^{\text{eff}}(t)$ for scenarios S1-S6.

Robustness analysis for μ

In Section 2.3.1, we address the identifiability challenges inherent in applying the SIPRO model, with a specific focus on the parameter μ. We emphasize the necessity of fixing μ to enhance the estimation of other crucial parameters. While fixing μ to the value of the simulated dataset during a simulation study is optimal, practical considerations arise when dealing with real-world data. Despite numerous clinical studies offering reliable estimates of the mean time to recovery for asymptomatic individuals (i.e., μ), we explored the robustness of our method to different choices of μ. Using the dataset simulated under scenario S1, with $\mu = 1/15$, we conduct a sensitivity analysis by running the algorithm nine times. We vary the choice of μ from a set of values $\{1/5, 1/10, 1/13, 1/14, 1/15, 1/16, 1/17, 1/20, 1/25\}$, which includes the correct value. For each run, we assess the goodness of fit by calculating the percentage of correctly estimated individual parameters, presented in the heatmaps in Figure 2.10. The top heatmap addresses trajectories initial values, α, and ν parameters, revealing that all parameter estimates, except for α, remain largely unaffected by changes in μ. As anticipated, for α, the estimation improves when μ is closer to the true value. The bottom heatmap pertains to time-dependent parameters: trajectories, $\beta_i(t)$, and $\rho_i^{\text{eff}}(t)$. The darker central portion of the map indicates that smaller differences between the true μ and the one used in estimation lead to improved method performance. In both cases, slight variations from the true μ value, such as $1/17 \leq \mu \leq 1/13$ in our sensitivity analysis, ensure consistently good method performance. Upon examining the percentage of parameter estimations with μ estimated and μ fixed to unrealistic values, the findings suggest a practical guideline for real-world applications. When there is the option to select a suitable value from existing studies, fixing μ leads to robust and reliable results. On the contrary, when reliable estimations for μ are not available in the literature, it is advisable to keep μ as a parameter to be estimated using the MCMC procedure. This approach allows flexibility in handling situations where precise information on μ is lacking, ensuring a more adaptive and accurate modeling of the data.

2.3.2 SIR statistical model to analyze Italian data: a simulation study

To compare the performances of our model with the simpler SIR, we also construct a statistical mixed model based on SIR dynamic and estimate the parameters of interest with a Bayesian structure. The epidemic evolution in the i-th region is described by an independent SIR-$(\beta_i(t), \mu_i)$ model, assuming no contacts can happen between the populations in different regions. We assume the data $\{y_{I,i}(t), y_{R,i}(t)\}$, realization of the ranodm variables $\{Y_{I,i}(t), Y_{R,i}(t)\}$, are noisy observations of the proportion of individuals in compartment I and R, for each region i-th region evaluated at day t. We assume a noise structure quite similar to Sub-section 2.3.1, and therefore we introduce a third complementary component $Y_{C,i}(t) = 1 - Y_{I,i}(t) - Y_{R,i}(t)$ which enters the random vector $\mathbf{Y} = \{Y_{I,i}(t), Y_{R,i}(t), Y_{C,i}(t)\}$. We model $\mathbf{Y}$ using a Dirichlet distribution centered at $(\imath_i(t), r_i(t), s_i(t))$ with a shared parameter γ that rules the noise amplitude (see Equation (2.27)). Italian regions heterogeneity is modeled through random effects on $(\beta_i(t), \mu_i)$: μ_i are drawn for a common lognormal population (with parameters μ and ω_μ, as in Equation (2.30)), while $\beta_i(t)$ are regional linear splines with random effect on parameters $\beta_{ik} = \beta_i(t_k)$ at prespecified equispaced days $\{t_k\}_{k=1}^K = (k-1)\Delta$, where Δ is the number of days between each node of the time grid and the following one. We impose that $\beta_i(t)$ is constant in the last time interval of the grid, namely between day t_{K-1} and day $t_K = T - 1$,

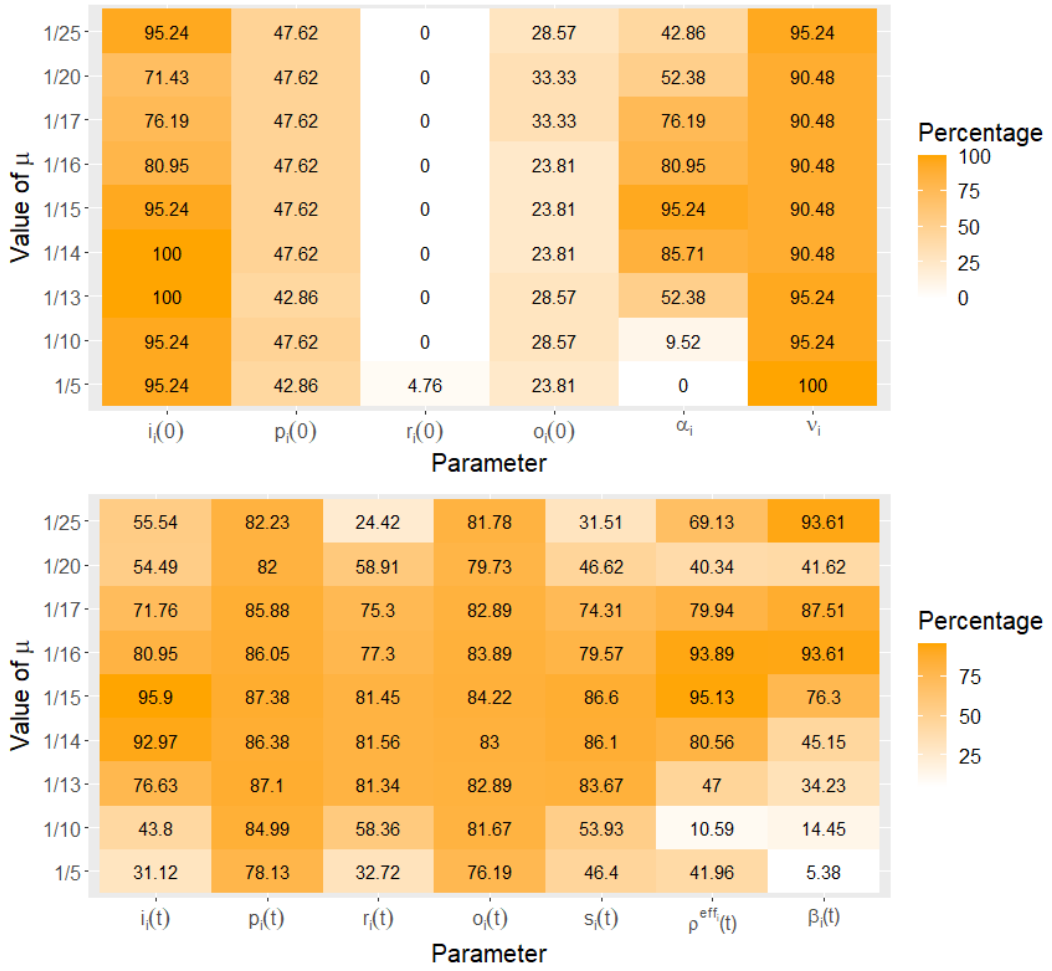

FIGURE 2.10: Heatmap with the percentage of individual parameters correctly estimated, darker colors indicate higher percentages. The estimations are obtained running the code for the same synthetic dataset (with μ=1/15) but fixing the value of μ to $\{1/5, 1/10, 1/13, 1/14, 1/15, 1/16, 1/17, 1/20, 1/25\}$.

for further details the reader is referred to the Sub-section 2.3.1. Initial conditions $x_i(0) = (s_i(0), \imath_i(0), r_i(0))$ are parameters that have to be estimated as well. We define the vector of *individual* (regional) parameters as

$$\boldsymbol{\phi} = \left(\{x_i(0)\}_{i=1,\dots,n}, \{\mu_i\}_{i=1,\dots,n}, \{\beta_{ik}\}_{\substack{i=1,\dots,n \\ k=1,\dots,K}} \right), \tag{2.24}$$

and the vector of the *global* parameters as

$$\boldsymbol{\theta} = \left(\log \mu, \omega_\mu, \gamma, \{\log \beta(k)\}_{k=1,\dots,K}, \omega_\beta \right). \tag{2.25}$$

The joint probability density of the model is

$$f(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{t=0}^{T} f(y_{I,i}(t), y_{R,i}(t), y_{C,i}(t) | \phi_i, \theta) f(\phi_i | \theta) f(\theta), \tag{2.26}$$

where $\boldsymbol{y} = \{y_{I,i}(t), y_{R,i}(t), y_{C,i}(t)\}_{\substack{i=1,\dots,n \\ t=0,\dots,T}}$, with $y_{C,i}(j) = 1 - y_{I,i}(j) - y_{R,i}(j)$, for all

i and t, $\phi_i = (x_i(0), \mu, \{\beta_{ik}\}_{k=1,...,K})$, and the conditional probability densities in the right-hand side are specified as follows:

$$Y_{I,i}(t), Y_{R,i}(t), Y_{C,i}(t)|\phi_i, \theta \overset{i.i.d.}{\sim} Dirichlet\Big(\gamma\, p_i(t),\, \gamma\, o_i(t),\, \gamma\,(1-\iota_i(t)-r_i(t))\Big), \quad (2.27)$$

$$\log(\beta_{ik})|\theta \overset{i.i.d}{\sim} \mathcal{N}(\log(\beta_k), \omega_\beta^2), \quad i = 1, ..., n,\ k = 1, ..., K, \quad (2.28)$$

$$\log(\mu_i)|\theta \overset{i.i.d}{\sim} \mathcal{N}(\log(\mu), \omega_\mu^2), \quad i = 1, ..., n, \quad (2.29)$$

$$(s_i(0), \iota_i(0), r_i(0)) \overset{i.i.d}{\sim} Dirichlet(8, 1, 1). \quad (2.30)$$

The prior distribution $f(\theta)$ is assumed to factorize as

$$f(\theta) = f(\mu)f(\omega_\mu)f(\gamma)f(\omega_\beta)\prod_{k=1}^{K} f(\beta(k)), \quad (2.31)$$

where

$$\log\beta(k) \sim \log\gamma \sim \log\mu \sim \mathcal{N}(0, 1000), \quad (2.32)$$

$$\omega_\beta^2 \sim \omega_\mu^2 \sim \mathcal{IG}(0.1, 0.1). \quad (2.33)$$

The statistical model presented is fully identifiable, as the SIR model is composed of only observed compartments. To sum up the performance of the MCMC algorithm we tested it on a simulated dataset. As we do for the SIPRO model in Section 2.3.1, we present the percentage of individual parameters correctly estimated (true values is in the 95% credible interval) in Table 2.8, the posterior 95% credible interval with the true value of global parameters in Table 2.9, and the adapted-MSE plots (see Definition 2.3.1) in Figure 2.11 to quantify the goodness of fit of the SIR trajectories. The overall performance is very good.

TABLE 2.8: Simulated SIR dataset - Results on individual parameters.

Parameter	Percentage
$\iota_i(0)$	95.24%
$r_i(0)$	95.24%
μ_i	90.48%
$\beta_i(t)$	90.48%
$\rho_i^{\text{eff}}(t)$	90.48%

TABLE 2.9: Simulated SIR dataset - Results on global parameters.

Parameter	2.5%	Mean	97.5%	Real
$\log(\nu)$	-3.011	-2.818	-2.614	-2.708
ω_ν	0.338	0.449	0.632	0.500
γ	5.937	5.989	6.035	6.000
ω_β	0.110	0.135	0.170	0.100
$\log(\beta_1)$	-0.408	-0.338	-0.272	-0.357
$\log(\beta_2)$	-1.707	-1.644	-1.584	-1.609
$\log(\beta_3)$	-1.286	-1.224	-1.163	-1.204
$\log(\beta_4) = \log(\beta_5)$	-1.262	-1.186	-1.115	-1.204

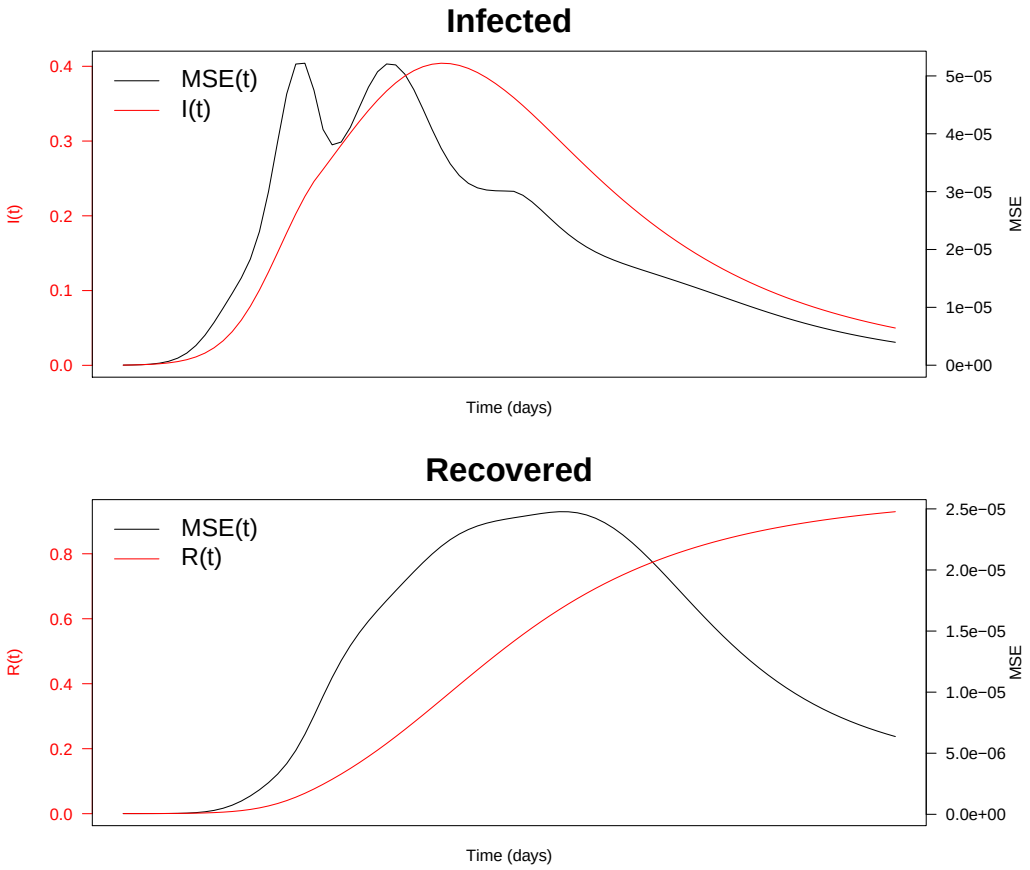

FIGURE 2.11: Mean MSE index (black) for each time to evaluate the fit of SIR-mixed-model, for Infected and Recovered trajectories (see Definition (2.3.1)). The red line is the number of Infected, Positive, Recovered, and Out in Italy for each day. The y-axes values of the red line are shown on the left-side, while the ones of the black lines are on right-side.

2.3.3 Real Data analyses and results

The results obtained on synthetic data, as presented in Sections 2.3.1 and 2.3.2, offer insights into the strengths and weaknesses of the SIR and SIPRO models, along with their respective MCMC estimation procedures. Both methods exhibit outstanding performance when applied to data generated from their respective models. Specifically, the SIR model, which encompasses all observed compartments, does not encounter any identifiability issues. On the other hand, the SIPRO model, incorporating both observed and unobserved compartments, may face practical identifiability challenges, as illustrated in 2.3.1. The motivation behind extending from the SIR to the SIPRO model is to achieve a more accurate representation that mimics the underlying mechanism of the COVID-19 pandemic. The ultimate goal of this work is to analyze public COVID-19 data and assess whether the developed SIPRO model can indeed improve the results obtained with the SIR model. Therefore, based on the simulation results, before applying the method to real data, careful consideration must be given to the characteristics of the data under analysis. This is essential to understand the applicability of the model and the reliability of the final results in a real-world context. We report, in Figure 2.12, the data collected from the Italian Protezione Civile [DPCa] from February 2020 to July 2021. The figure shows the overall evolution of the pandemic in Italy, each color represents a different region (21 regions are reported as autonomous regions are counted as well). The plot reveals a major phase that characterized the pandemic, followed by a period of consistently low cases during the summer of 2020. In particular, as per the discussion on the weak identifiability problems highlighted in the previous section, to be able to model the data we should divide the first wave data from the low-case period in Summer 2020. We choose the set one node every 3 weeks, to capture weekly patterns and to avoid having to estimate a huge number of parameters. We estimate the parameters of interest with SIPRO and SIR mixed-effect models.

To compare the goodness of fit with SIR and SIPRO mixed models (with fixed $1/\mu = 1, 3, 5, 7, 10$, for the reason explained in Sub-section 2.3.1) to the selected data, we report the WAIC index (see Table 2.10), as defined in [GHV14]. The performances are quite similar overall, but SIPRO with $\mu = 1/5$ was selected to give the best fit (outperforming also the SIR model). All the further analyses involving the SIPRO model are performed assuming $\mu = 1/5$.

To give an overall idea of the fit of SIPRO estimates, we randomly selected three regions, Abruzzo, Toscana, and Lombardia, and report the estimated trajectory, with their 95% credible interval, in Figure 2.13: the SIPRO model gives a good description

TABLE 2.10: Comparison of SIPRO ($1/\mu$ = 3, 5, 7, 10) and SIR model performances with WAIC and ES indexes. For both indexes, the lower (in bold) is the best.

Method (First Period)	WAIC	ES_1	ES_2
SIR	$-254.09 \cdot 10^4$	$56.12 \cdot 10^{-4}$	$68.05 \cdot 10^{-4}$
SIPRO ($1/\mu = 3$)	$-246.43 \cdot 10^4$	$\mathbf{42.58 \cdot 10^{-4}}$	$69.84 \cdot 10^{-4}$
SIPRO ($1/\mu = 5$)	$\mathbf{-255.51 \cdot 10^4}$	$43.12 \cdot 10^{-4}$	$69.60 \cdot 10^{-4}$
SIPRO ($1/\mu = 7$)	$-255.03 \cdot 10^4$	$43.23 \cdot 10^{-4}$	$69.54 \cdot 10^{-4}$
SIPRO ($1/\mu = 10$)	$-254.11 \cdot 10^4$	$43.04 \cdot 10^{-4}$	$\mathbf{67.11 \cdot 10^{-4}}$

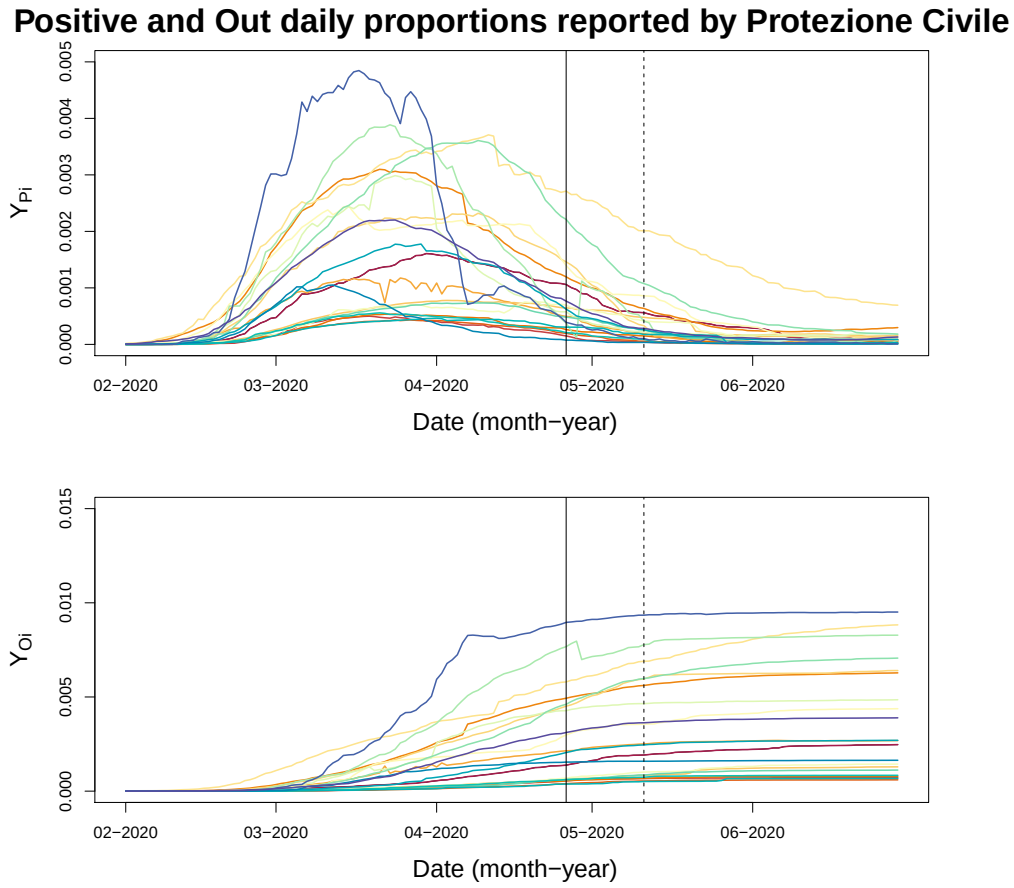

FIGURE 2.12: Daily proportion reported by the Italian Protezione Civile for the 21 considered regions [DPCa]: different colors stand for different regions. Black solid line is the end of the time window considered for estimation, black-dotted line is the end of the time window considered for prediction.

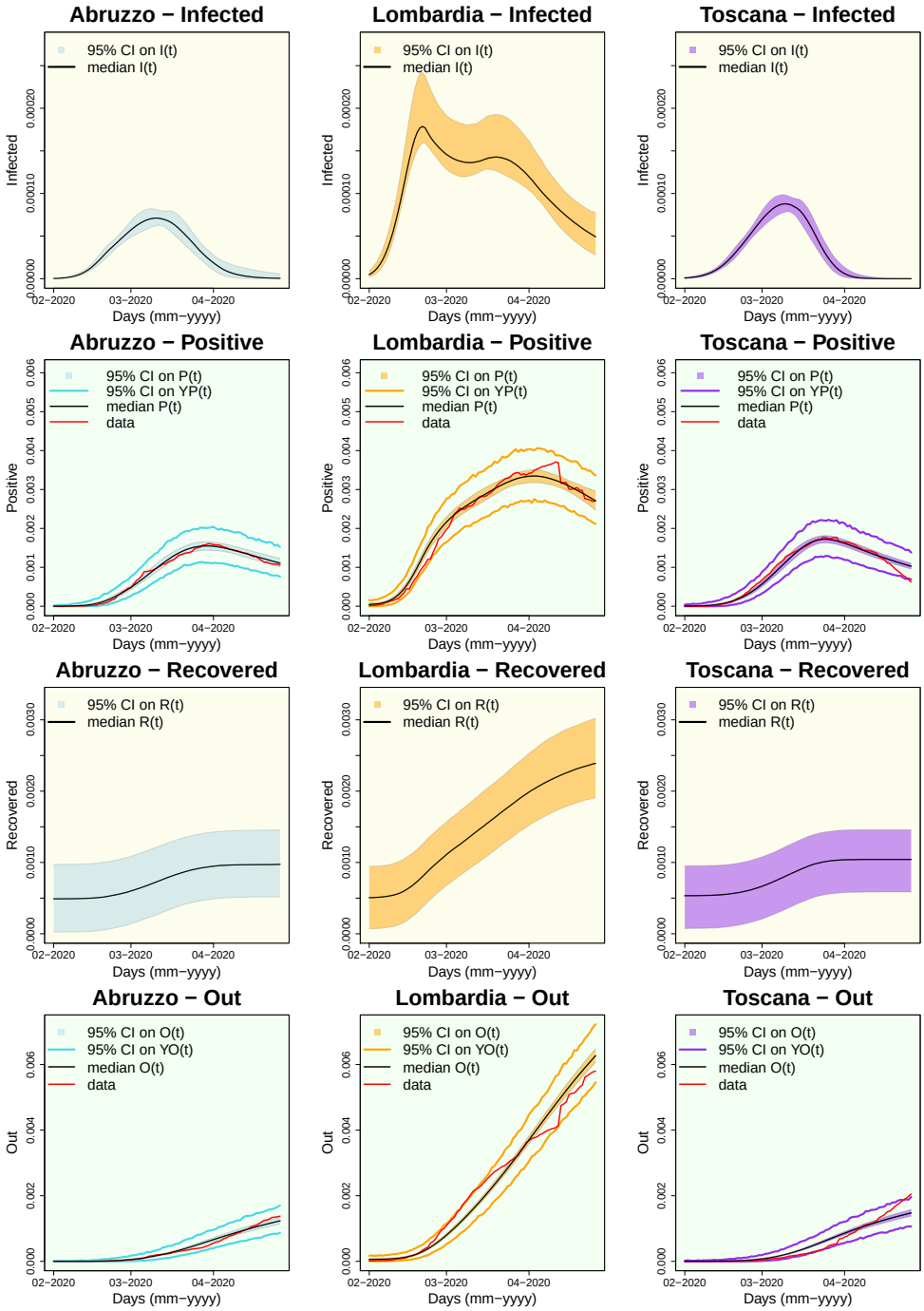

FIGURE 2.13: Median estimation, with 95% credible of Infected, Positive, Recovered and Out, obtained with SIPRO ($1/\mu = 5$). Predictive interval 95% is estimated for Positive and Out. Yellow background stands for unobserved compartments of the model, and green for observed compartments. Data are represented by red lines.

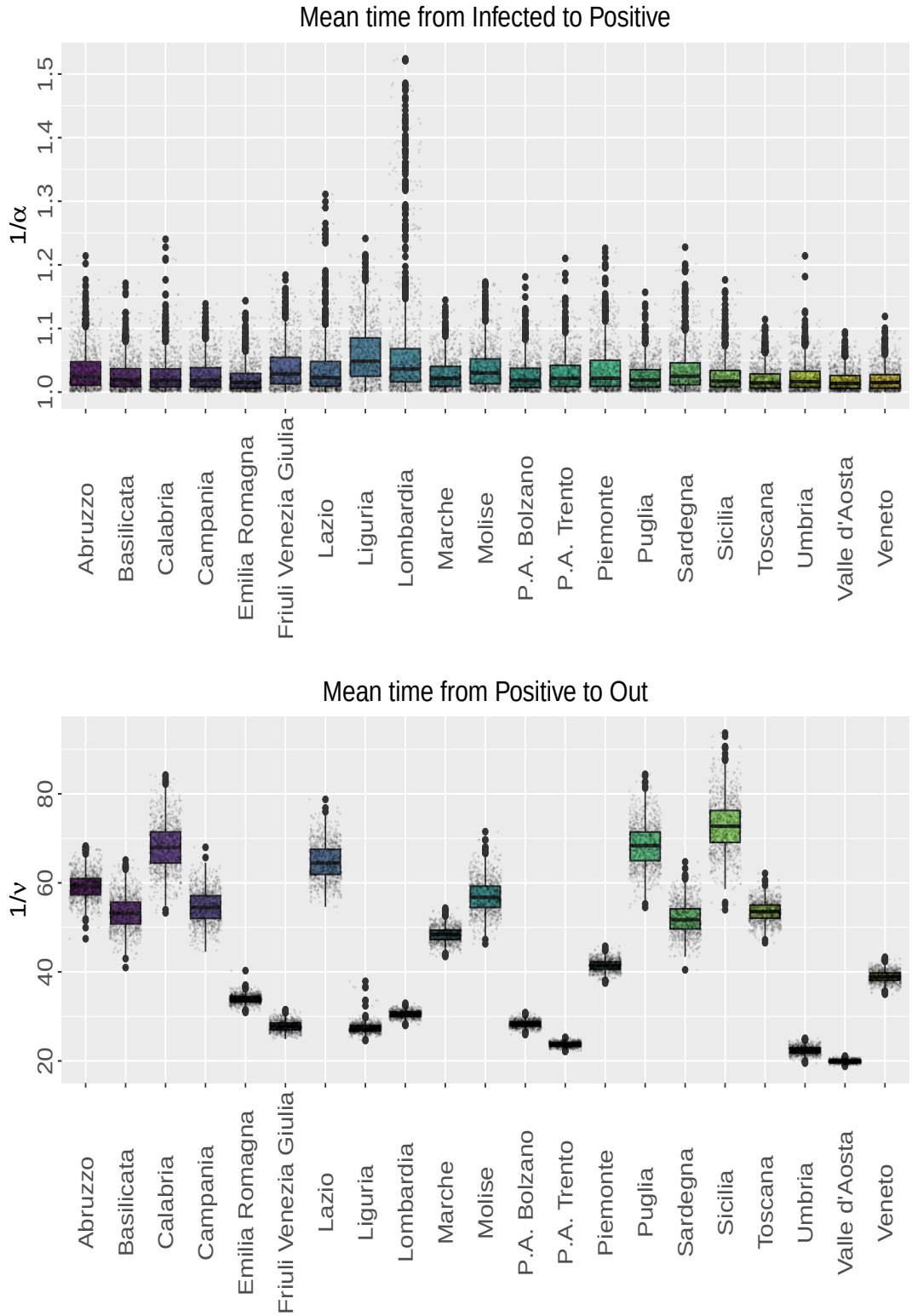

FIGURE 2.14: On the top: posterior boxplot of $1/\alpha_i$ with posterior samples (points). On the bottom: posterior boxplot of $1/\nu_i$ with posterior samples (points).

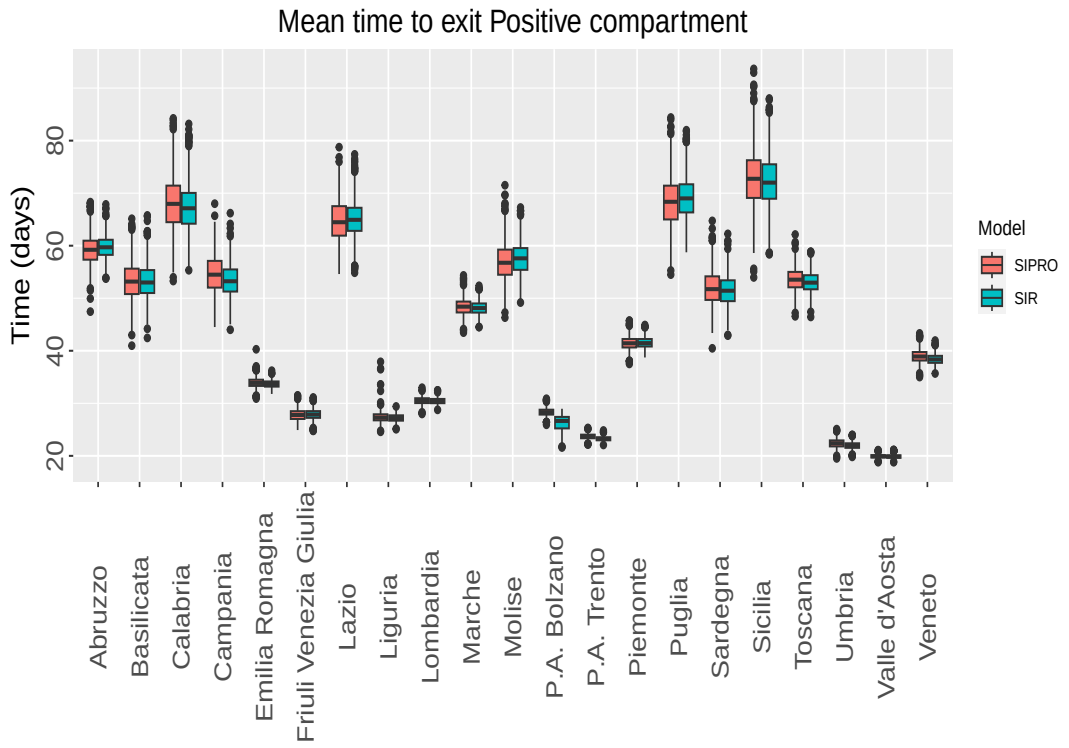

FIGURE 2.15: Posterior boxplot to compare mean time of transition between observed compartments in SIR and SIPRO model: the parameter of interest is $1/\mu_i$ in the SIR model and $1/\nu_i$ in the SIPRO model, but they refer to the same quantity.

of the observed data, even if the underlying parameters are not always perfectly estimated. In particular, we report the posterior distribution of the mean time of the transition from Infected to Positive $1/\alpha_i$ and from Positive and Out $1/\nu_i$ in Figure 2.14: the first quantity, which is close to one for each region, is a little bit lower than we expect, and is probably suffering of the low-identifiability problems we already discussed when we analyzed the SIPRO simulations (Sub-section 2.3.1). The second one, which estimates are absolutely reliable, reflects the delay in reporting the information that characterized the first wave of the pandemic. To confirm the latter result, we compare the estimated transition time between observed compartments, with both SIR and SIPRO dynamics, in Figure 2.15.

To better motivate the necessity of extending the SIR model to a more complex one, we compare the SIR and SIPRO (with $1/\mu = 5$) in terms of reproductive number estimation and new case predictions.

Comparison of basic reproduction numbers

The performances of SIR and SIPRO models in fitting the data and the evolution of the pandemic are very good and quite similar, despite some of the common underlying parameters being hugely different: in particular, the estimated basic and effective reproduction number are completely different, as they assume a different source of contagiousness. With SIR and SIPRO mixed model we estimated the regional basic and effective reproduction number for each region, but to graphically show their difference we only report, in Figure 2.16, a national estimation of $\rho^{\text{eff}}(t)$,

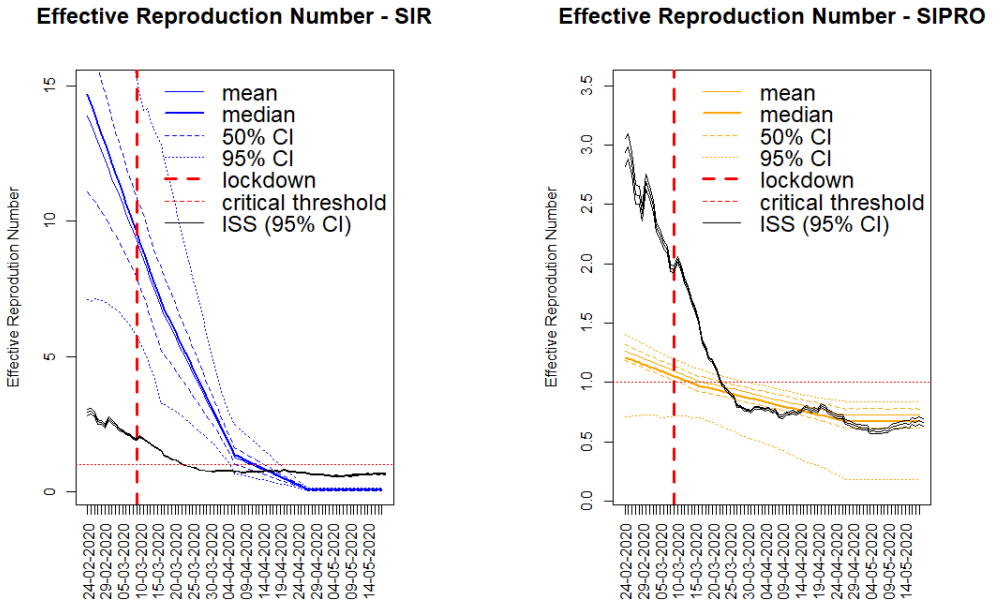

FIGURE 2.16: National effective reproduction number on first phase, day by day (dd-mm-yyyy), estimated with SIR mixed model (on the left, in blue), SIPRO ($\mu = 1/5$) mixed model (on the right, in orange), and with ISS code in black [GM20]. The vertical red line indicates the starting date of the lockdown, while the horizontal red line indicates the epidemic spread threshold, set to 1.

computed from the posterior B sample as

$$\rho^{\text{eff}_{\text{SIPRO}},b}(t) = \frac{\beta^b(t)S^b(t)}{\mu + \bar{\alpha}^b} \quad \text{and} \quad \rho^{\text{eff}_{\text{SIR}},b}(t) = \frac{\beta^b(t)S^b(t)}{\bar{\alpha}^b}, \quad B = 1, ..., b, \tag{2.34}$$

where $\beta^b(t)$ is the linear spline interpolating the mean of random effects $\{\beta_k^b\}_{k=1,...,K}$, $S^b(t)$ the national proportion of Susceptibles, $\bar{\alpha}^b$ the mean of individual parameters $\{\alpha_i^b\}_{i=1,...,n}$, and μ fixed to $1/5$.

Not surprisingly, $\rho^{\text{eff}_{\text{SIR}}}(t)$ is ten times higher than $\rho^{\text{eff}_{\text{SIPRO}}}(t)$, as it does not account for asymptomatics and, thus, explains all the registered cases to arise from previously registered infections.

We also compare $\rho^{\text{eff}_{\text{SIPRO}}}(t)$ and $\rho^{\text{eff}_{\text{SIR}}}(t)$ with the estimations provided by the Government in [GM20], which are based on different data sources. The two estimations, $\rho^{\text{eff}_{\text{SIR}}}(t)$ and $\rho^{ISS}(t)$, show significant differences. At the beginning of the epidemic, $\rho^{\text{eff}_{\text{SIR}}}(t)$ is more than three times greater than $\rho^{ISS}(t)$. Additionally, $\rho^{\text{eff}_{\text{SIR}}}(t)$ takes considerably longer to drop below the key threshold of one compared to $\rho^{ISS}(t)$. Moreover, towards the end of the pandemic, the SIR estimation approaches almost zero, which is unrealistic.

On the contrary, even though the SIPRO and ISS estimations differ noticeably during the first month of the pandemic, they exhibit similar behavior concerning the critical threshold. Both estimations are above the critical threshold, indicating a dramatic evolution of the epidemic during the initial phase and becoming quite similar afterward.

Evaluating the predictions with Energy Score

More than simply modeling the daily count of positive and infected, it is interesting to make short-term predictions. In particular, predictions can be helpful to promptly make healthcare and political decisions. To explain how to make predictions with SIPRO model we look at fifteen days predictions, after the last measured time-point on phase one. To better validate how good our model is in making good provisions, with respect to standard SIR, we fit SIPRO and SIR models and then we make predictions for fourteen days after the last observed data, keeping all the time-dependent parameters fixed to the value estimated at the final fitting time. To evaluate the predictions, and to avoid reporting all the estimates, we use the Energy Score, as described in [JKL19], and defined as

Definition 2.3.2 (Energy Score for SIPRO model)**.** *Let* $\mathbf{P}^b = \{P_i^b(t)\}_{t=T+1,\dots,T+14}^{i=1,\dots,n}$ *and* $\mathbf{O}^b = \{O_i^b(t)\}_{t=T+1,\dots,T+14}^{i=1,\dots,n}$ *be the b-th sampled proportion of positive and out for each region and for each day of the prediction interval, for the SIPRO model. We evaluate the goodness of our estimates by calculating the energy score for Observed Positive and Observed Out, respectively as*

$$ES_1 = \frac{1}{B}\sum_{b=1}^{B} \|\mathbf{P}^b - \mathbf{Y}_P\|_2 - \frac{1}{2B^2}\sum_{b=1}^{B}\sum_{m=1}^{B} \|\mathbf{P}^b - \mathbf{P}^m\|_2, \tag{2.35}$$

$$ES_2 = \frac{1}{B}\sum_{b=1}^{B} \|\mathbf{O}^b - \mathbf{Y}_O\|_2 - \frac{1}{2B^2}\sum_{b=1}^{B}\sum_{m=1}^{B} \|\mathbf{O}^b - \mathbf{O}^m\|_2, \tag{2.36}$$

with $\mathbf{Y}_P = \{Y_{P,i}(t)\}_{t=T+1,\dots,T+15}^{i=1,\dots,n}$ *and* $\mathbf{Y}_O = \{Y_{O,i}(t)\}_{t=T+1,\dots,T+15}^{i=1,\dots,n}$ *the true observed proportions we want to predict, and* $\|\cdot\|_2$ *the Euclidean norm.*

Definition 2.3.3 (Energy Score for SIR model)**.** *Let* $\mathbf{I}^b = \{I_i^b(t)\}_{t=T+1,\dots,T+14}^{i=1,\dots,n}$ *and* $\mathbf{R}^b = \{R_i^b(t)\}_{t=T+1,\dots,T+14}^{i=1,\dots,n}$ *be the analogous quantity for the SIR model, then the energy scores can be computed as*

$$ES_1 = \frac{1}{B}\sum_{b=1}^{B} \|\mathbf{I}^b - \mathbf{Y}_I\|_2 - \frac{1}{2B^2}\sum_{b=1}^{B}\sum_{m=1}^{B} \|\mathbf{I}^b - \mathbf{I}^m\|_2, \tag{2.37}$$

$$ES_2 = \frac{1}{B}\sum_{b=1}^{B} \|\mathbf{R}^b - \mathbf{Y}_R\|_2 - \frac{1}{2B^2}\sum_{b=1}^{B}\sum_{m=1}^{B} \|\mathbf{R}^b - \mathbf{R}^m\|_2, \tag{2.38}$$

with $\mathbf{Y}_I = \{Y_{I,i}(t)\}_{t=T+1,\dots,T+15}^{i=1,\dots,n}$ *and* $\mathbf{Y}_R = \{Y_{R,i}(t)\}_{t=T+1,\dots,T+15}^{i=1,\dots,n}$ *the true observed proportions we want to predict, and* $\|\cdot\|_2$ *the Euclidean norm.*

We report the evaluated energy scores in Table 2.10: the prediction of new Positive individuals is better performed by SIPRO model, while the prediction of new Recovered/Out individuals is better with the SIR. This result is consistent with the mechanism described by the models under analysis: the SIPRO model better describes the real dynamic of the pandemic involving symptomatic and asymptomatic individuals and thus gives better intuition of the virus transmission, on the other hand, the SIR model better describes the transition between observed compartments and, thus, gives a more reliable estimation of how tested people will recover. We only report the full prediction for the three selected regions, in Figure 2.17.

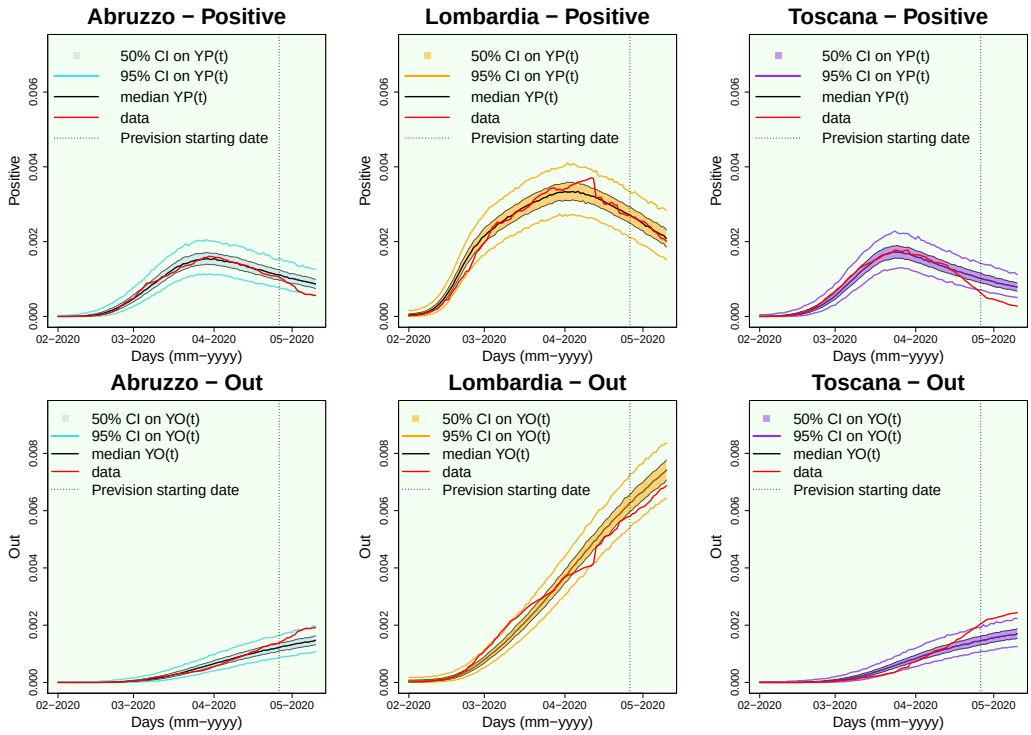

FIGURE 2.17: Median $Y_P(t)$ and $Y_O(t)$ estimated with SIPRO ($1/\mu = 5$), with 50% and 95% credible interval. Red line represents the data. Vertical black-dotted line divides the time points used for estimation and the time points predicted by the model.

2.4 Final remarks and conclusions

The aim of the SIPRO model is to improve upon the SIR model, which inadequately described the COVID-19 pandemic. The goal is to create a simple yet identifiable model that avoids an excessive number of arbitrarily chosen parameters. Furthermore, the main objective is to construct a realistic model capable of estimating the reproductive number, the daily number of infections, and the time from infection to positivity. We develop a statistical mixed model that combines the SIPRO dynamics with a random effect structure, enabling us to account for the heterogeneity among the Italian regions during the initial phase of the pandemic. The parameter estimation is performed using a Metropolis-within-Gibbs algorithm combined with Parallel tempering updates, which help us addressing low-identifiability issues. The model performs well in estimating trajectories (both observed and unobserved) and provides accurate short-term predictions, outperforming the SIR dynamic. Although the estimation of the reproductive number does not perfectly align with the ISS estimation, the model correctly identifies when this index is above or below the critical threshold of one, which is not possible with the SIR model. However, some identifiability problems persist with parameters β, α, and μ, representing the transitions from Susceptible to Infected, from Infected to Positive, and from Infected to Recovered, respectively. These parameters individually face low-identifiability issues, but they compensate for each other, resulting anyway in a good fit for the data.

Chapter 3

Estimating the optimal time to perform a PET-PSMA exam in prostatectomized patients based on data from clinical practice

This chapter is based on the paper:
Amongero, Martina and Mastrantonio, Gianluca, and De Luca, Stefano, and Gasparini, Mauro (2023). **Estimating the optimal time to perform a PET-PSMA exam in prostatectomized patients based on data from clinical practice.** (Under review). [Amo+23]

Nowadays, one of the most important areas of medical statistical development is oncology. The statistical and clinical interest varies from the early detection of tumor presence and locations, the estimation of (personalized) treatment efficacy, the definition of optimal personalized treatment strategies, the analyses of resurgences, the development of ethical trial structures, to many others. This chapter focuses on the early detection of locations of disease and arises from a collaboration with the clinicians of the Department of Urology of San Luigi Gonzaga Hospital in Torino (Italy).

The main focus of this work is to analyze data from an observational study on prostatectomized patients. Prostatectomized patients are at high risk of resurgence: this is the reason why, during a follow-up period, they are monitored for PSA (Prostate-Specific Antigen) growth, an indicator of tumor progression. The presence of tumors can be evaluated with an expensive exam, called PET-PSMA (Positron Emission Tomography with Prostate-Specific Membrane Antigen). To justify the high cost of the PET-PSMA and, at the same time, to contain the risk for the patient, this exam should be recommended only when the evidence of tumor progression is strong. To estimate the optimal time to recommend the exam based on the patient's history and collected data, we build a hierarchical Bayesian model that describes, jointly, the PSA growth curve and the probability of a positive PET-PSMA, basing it on a Gompert model. With our proposal we process all past and present information about the patients PSA measurement and PET-PSMA results, in order to give an informed estimate of the optimal time, improving current practice.

The chapter is divided as follows: before going into the details of our work, we give a small recap of the Gompertz model 3.1, and we sum up the literature for the problem of interest 3.2. Then, In Section 3.3, we present the joint model for PSA measurements and test results. Section 3.4 is devoted to the definition and the estimation

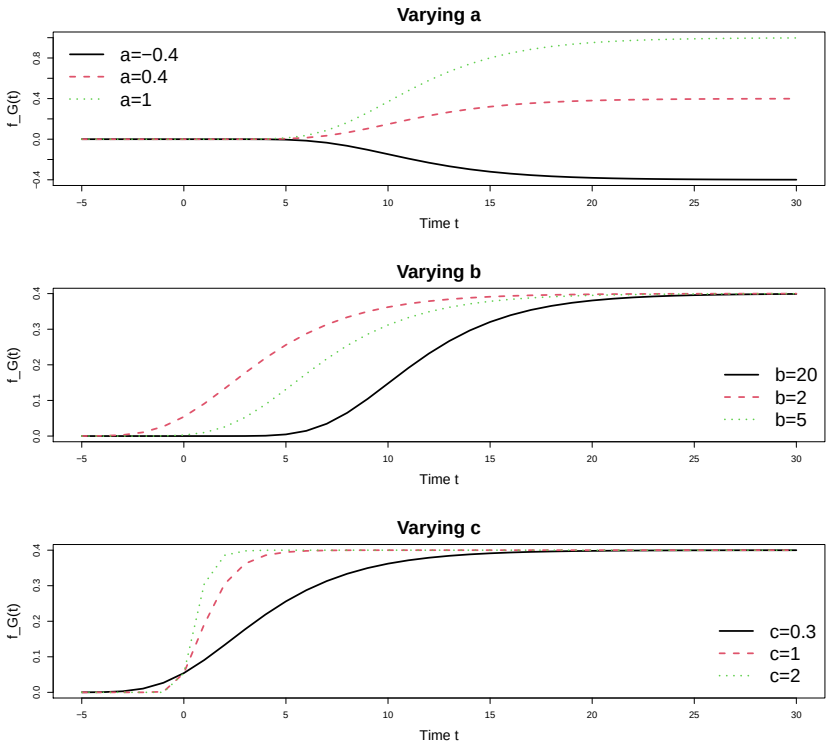

FIGURE 3.1: Gomperts functions $f_G(t)$ varying the parameters a, b, c.

of the optimal time to perform a test. We test our model on simulated datasets, as explained in Section 3.5, before applying it to real data, Section 3.6 contains the results of the model estimated on a group of patients from San Luigi Gonzaga Hospital, in Torino, Italy. The Chapter ends with some conclusive remarks in Section 3.7.

3.1 Gompertz model

The Gompertz curve is a famous function that can be used to describe time series, named after Benjamin Gompertz. It is a sigmoid function, with left asymptotic value (or lower-valued asymptote) equal to 0 and the right one (or future value asymptote) which is a parameter of the function. The right-side (or future value asymptote) of the function is approached much more gradually by the curve than the left-side (or lower-valued asymptote). It is closely related to the logistic function but, while the logistic function is anti-symmetrical with respect to its central point (the x-value where the central point between the two asymptotes is reached), the Gompertz curve is more flexible, allowing different rates of changes in reaching the asymptotes. The function is often applied in literature to describe tumor growth when physical barriers prevent the tumor to spread out [Xu87]. The equation depends on three different values a, b, and c, as

$$f_G(t) = ae^{-be^{-ct}}, \tag{3.1}$$

where $a = \lim_{t\to+\infty} f_G(t)$, b displaces the graph on x-axis, and finally c gives the growth rate. In Figure 3.1 some Gompertz functions are plotted to show which roles play the parameters a, b, and c on the shape of the function.

3.2 Literature

Prostate cancer is the most frequent neoplasm in men, with an incidence of around 7% among all new cancer cases [IAR], and has therefore attracted a lot of interest in the last twenty years. Research focuses on very different topics, starting from the causes and the incidence of the tumor, modeling its growth, analyzing the effect of therapies, and, finally, analyzing the risk, the locations, and time of biochemical relapse (BCR) and clinical recurrences.

High levels of serum Prostate-Specific Antigen (PSA) after primary treatments, such as surgery or radiotherapy, were identified in the literature to be significant indicators of tumor progression taking place at some locations different from the original one [Afs+17; Eib+16; Ver+16; Fen+19; Hof+19]. This means that prostatectomized patients are at risk of developing BCR and metastasis, which is the reason why, during a follow-up period of several years, they are usually monitored for PSA recurrence.

PET-PSMA is a nuclear medicine survey, that is currently one of the most sensitive tests for the early detection of tumor presence and location. It is very expensive and complex and, for this reason, patients should be referred to a PET-PSMA exam only in case of confirmed high risk of BCR. The timing of the exam is very important since if it is performed too late, the patient is subject to excessive risks, while, if it is too soon, there is a high probability of false negative results. The estimation of the optimal time to perform such an examination is still an open problem widely discussed in the literature.

Currently, the most used indicator of BCR is the PSA Doubling Time (PSA-DT) which is usually monitored over time to control the PSA evolution [Reg+20]. This indicator is calculated using only the last two measurements available for the patient analyzed. According to literature [Reg+20], a PSA-DT<6 months, together with a thorough clinical examination of the patient stage and conditions, should be used as an indication to perform a PET-PSMA exam. Some authors model the PSA evolution over time trying to link it to tumor progression [SC97; Hir+12], distinguishing between patients with and without disease [Car+92; PT09], or accounting of the impact of aging on its evolution [Pea+91]. Other work mainly focuses on the PET examination [Fos+19] or on the correlation between PSA levels and PET results [Per+19]. Finally, some researchers link PSA kinetics and tumor recurrence [PT09], whereas other authors try to determine the optimal time of PET-PSMA [Lui+20] regardless of the kinetics of PSA. Our work combines many aspects of all these papers into an overall joint statistical model [VM97; TD04; DJ08], and exploits its structure to make predictions. We base our model construction on the joint-model idea proposed by [PT09]. However, our final aim differs: while they jointly model the PSA growth curve and the probability of a resurgence, we substitute for the latter part a logistic model for the probability of positive PET-PSMA results in the presence of resurgences. Specifically, we focus on estimating more precisely the evolution of the probability of a positive PET-PSMA and, consequently, determining the optimal time to recommend such an exam in cases of suspected BCR. The aim of the present study is to gain accuracy by basing decisions on the individual clinical and serological patient history and on an entire database of similar patients, built on clinical practice, rather than relying solely on the last two individual measurements of PSA. To do so, a Bayesian Network [CP12] is built, or more precisely a Bayesian hierarchical model. Bayesian Networks are Probabilistic Graphical Models which represent a set of variables and their probabilistic interdependencies, allowing for the data analyst to have

an accurate uncertainty quantification of all model unknown parameters.

We assume that the true, yet unobservable, PSA levels depend on patient covariates and on the presence of BCR, the latter being a model parameter, which is then estimated during the model fitting. Specifically, as modeled by [PT09], the latent PSA level is expected to decrease before BCR and to increase afterward. Both the measured PSA level and PET-PSMA results are dependent on this latent structure, which is probabilistically linked with them. One of the significant advantages of our approach is that we are utilizing all available information jointly to determine the presence of BCR and the likelihood of a positive PET-PSMA. Moreover, in contrast to other methods that link PET-PSMA results to the measured PSA level, our approach allows for estimating the probability of a positive PET-PSMA without an associated PSA measurement. By selecting a target probability of having a positive PET-PSMA and a confidence level, we can estimate the optimal time to perform the test with Markov Chain Monte Carlo methods. Inference is performed according to the principles of Bayesian Statistics, which allows us to have a complete evaluation of the uncertainty associated with each model component.

3.3 Joint model of PSA growth curve and PET-PSMA results

This Section is devoted to explain the statistical joint model: we first describe the growth model for the PSA curve (Sub-section 3.3.1), and then how this model is linked to the logistic structure for the result of PET-PSMA (Sub-section 3.3.2). Finally, Sub-section 3.3.3 deals with random effects.

3.3.1 The data and its likelihood

Let $x_i(t)$ be the PSA level of the i-th patient at time t, where $t = 0$ indicates the time the patient has had a prostatectomy (only operated patients are considered in this work) and $i = 1, 2, \ldots, n$, where n is the total number of patients in the study. The recorded variable $y_i(t)$ is a noisy version of the non-negative quantity $x_i(t)$, therefore we assume

$$\log y_i(t) \sim \mathcal{N}(\log x_i(t), \sigma_i^2). \tag{3.2}$$

The choice of the functional form of $x_i(t)$, i.e., how PSA levels depend on time and other covariates, is one of the two major components of our proposed joint model (described in Sub-section 3.3.2), but before we discuss it, let us introduce the second component, which is the probability $\pi_i(t)$ for patient i at time t that, if a PET-PSMA is taken, the result will be positive. For each patient, the test outcome $z_i(t) \in \{0, 1\}$ is assumed to be Bernoulli random variable with a patient-specific and time-dependent probability, i.e.,

$$z_i(t) \sim Ber(\pi_i(t)). \tag{3.3}$$

Variables $x_i(t)$ and $\pi_i(t)$ are not observed, and we can only observe $y_i(t)$ and $z_i(t)$ at specific time points. The time points where these two variables are recorded can differ within and between patients, and we indicate as $\mathcal{T}_{y,i}$ and $\mathcal{T}_{z,i}$ the set of time points of patient i where $y_i(t)$ and $z_i(t)$ are recorded, respectively. The set $\mathcal{T}_i = \mathcal{T}_{y,i} \cup \mathcal{T}_{z,i}$ is the set of points where we have a measure of at least one of the two variables.

Let $\mathbf{y}_i^o = \{y_i(t)\}_{t \in \mathcal{T}_{y,i}}$, $\mathbf{x}_i^o = \{x_i(t)\}_{t \in \mathcal{T}_{y,i}}$ be the observed and latent PSA values over the time points in $\mathcal{T}_{y,i}$ and let $\mathbf{z}_i^o = \{z_i(t)\}_{t \in \mathcal{T}_{z,i}}$, $\boldsymbol{\pi}_i^o = \{\pi_i(t)\}_{t \in \mathcal{T}_{z,i}}$ be the test results and probabilities over $\mathcal{T}_{z,i}$ (i.e., multiple tests for the same patient can be

collected, at multiple times, that do not necessarily correspond to times PSA measurements are collected). Conversely, using the superscript u for "unobserved" (as opposed to the previous o for "observed"), let $\mathbf{y}_i^u = \{y_i(t)\}_{t\in\mathcal{T}_{z,i}}$, $\mathbf{x}_i^u = \{x_i(t)\}_{t\in\mathcal{T}_{z,i}}$, $\mathbf{z}_i^u = \{z_i(t)\}_{t\in\mathcal{T}_{y,i}}$, $\boldsymbol{\pi}_i^u = \{\pi_i(t)\}_{t\in\mathcal{T}_{y,i}}$ the vectors of variables at the time points where the associated process is not measured, i.e., $(y_i(t), x_i(t))$ at time points $\mathcal{T}_{z,i}$ where the exams are taken, and $(z_i(t), \pi_i(t))$ at time points $\mathcal{T}_{y,i}$ where the PSA is measured, respectively.

High values of PSA are indicators of tumor progression and are thus associated with a high probability of positive PET-PSMA results. We will then assume that $\pi_i(t)$ is a function of $x_i(t)$, and use this relation to find the optimal PET-PSMA time since the central idea of this work is to exploit the joint model structure [TD04; PT09] to describe the available longitudinal data. To specify the model we have to define the joint density of variables $(\log y_i(t), z_i(t))'$ over $\mathcal{T}_i$, for both observed and missing data. Hence, the missing data are considered further parameters to be estimated during the model fitting, which are easy to estimate within the Bayesian framework.

The joint density for a random sample is then factorized in the following way

$$\begin{aligned} &f\big(\log \mathbf{y}^u, \log \mathbf{y}^o, \mathbf{z}^u, \mathbf{z}^o \mid \log \mathbf{x}^u, \log \mathbf{x}^o, \boldsymbol{\pi}^u, \boldsymbol{\pi}^o, \boldsymbol{\theta}\big) = \\ &\prod_{i=1}^{n}\prod_{t\in\mathcal{T}_i} f\big(\log y_i(t) \mid \log x_i(t), \boldsymbol{\theta}\big) f\big(z_i(t) \mid \log x_i(t), \pi_i(t), \boldsymbol{\theta}\big), \end{aligned} \tag{3.4}$$

where $f(\cdot)$ stands for a generic probability density function (to be identified by its arguments) and $\boldsymbol{\theta}$ is a vector of parameters. In Equation (3.4) we are assuming that $(\log y_i(t), z_i(t))$ are conditionally independent given the latent variables since the connection between the two measurements $y_i(t)$ and $z_i(t)$ is through the latent PSA level $x_i(t)$, and once we know $x_i(t)$, no further information is needed to explain the PET-PSMA result. In the next paragraphs we illustrate our proposals for $f(\log y_i(t) \mid \log x_i(t), \boldsymbol{\theta})$ and $f(z_i(t) \mid \log x_i(t), \pi_i(t), \boldsymbol{\theta})$. We want to remark that in Equation (3.4) we are assuming conditional independence between the measurements, but, on the other hand, as we show in Sub-section 3.3.3, we introduce random effects over some components of $\boldsymbol{\theta}$, to model a form of dependence.

3.3.2 Joint model for each patient

To model the time evolution of PSA levels, we assume that it is composed of two phases, the first one, right after prostatectomy, where the PSA level is stationary or even decreasing over time, and the second one, after a patient-dependent change point in time τ_i, where it is assumed that the PSA increases, until reaching a plateau after some time, that we indicate as a_i. For each patient, the time τ_i can be interpreted as the unknown time at which resurgence starts, which is a crucial object of inference in our model. The component $\log x_i(t)$ is modeled as a linear function with patient-specific intercept, λ_i and slope $-\mu_i$, for non negative μ_i, i.e., if $t \leq \tau_i$

$$\log x_i(t) = \lambda_i - \mu_i t, \tag{3.5}$$

while for $t > \tau_i$, we model $\log x_i(t_i)$ as a weighted mean of the value assumed by $\log x_i(t)$ at the change point, i.e., $\log x_i(\tau_i)$, and its asymptotic value a_i:

$$\log x_i(t) = \log x_i(\tau_i) e_i(t) + a_i(1 - e_i(t)). \tag{3.6}$$

The weight function $e_i(t)$ is defined as

$$e_i(t) = \exp(-\gamma_i(t - \tau_i)), \tag{3.7}$$

with $\gamma_i \in \mathbb{R}^+$, to ensure that $\log x_i(t)$ is continuous at the change point τ_i, and it reaches the asymptotic value a_i as $t \to \infty$. Hence, for $t > \tau_i$, $x_i(t)$ is a version of the log-Gompertz growth function (3.1) with rate γ_i. We assume non-negative μ_i and γ_i to model the decreasing phase before τ_i and the increasing phase after it.

The second part of the joint model is instead focused on the binary PET-PSMA results $z_i(t)$, and in particular on modeling its probability $\pi_i(t)$. To connect the probability $\pi_i(t)$ to the PSA levels we use a logistic regression model

$$\text{logit}\{\pi_i(t)\} = \beta_{0,i} + \beta_1 \log x_i(t) + \beta_2 t, \tag{3.8}$$

where the linear prediction has a patient-specific intercept and an extra term linear on time to model temporal trends that are not be explained by $\log x_i(t)$. It should be noted that we define the relation in terms of the true latent PSA levels $x_i(t)$, not the observed ones $y_i(t)$. We expect β_1 to be positive since the larger the PSA, the larger the probability of a positive test. It is easy to see, from Equation (3.8), that, if $\beta_1 > 0$, then $\pi_i(t)$ goes to zero as $x_i(t)$ goes to zero:

$$\lim_{x_i(t)\to 0^+} \pi_i(t) = \lim_{x_i(t)\to 0^+} \frac{x_i^{\beta_1}(t) e^{\beta_{0,i}+\beta_2 t}}{1 + x_i^{\beta_1}(t) e^{\beta_{0,i}+\beta_2 t}} = 0. \tag{3.9}$$

All the unknown quantities in the model $\{(\lambda_i, \mu_i, a_i, \gamma_i, \tau_i, \beta_{0,i}, \beta_1, \beta_2, \sigma_i^2)\}_{i=1,\dots,n}$ make up the parameter vector $\boldsymbol{\theta}$.

3.3.3 Random effects

We now extend the base model, where each patient has its own set of parameters, to a random effects model. Generally, random effects can help to account for variability and heterogeneity in the data, due to unobserved or unmeasured factors, and may lead to more accurate and reliable estimates of treatment effects.

More precisely, we assume that the model describing PSA evolution can be enriched by the following second-level distributions:

$$\begin{aligned} \log \mu_i &\sim \mathcal{N}(\psi_{i,\mu}, w_\mu^2), \\ \log \gamma_i &\sim \mathcal{N}(\psi_{i,\gamma}, w_\gamma^2), \\ a_i &\sim \mathcal{N}(\psi_{i,a}, w_a^2), \\ \lambda_i &\sim \mathcal{N}(\psi_{i,\lambda}, w_\lambda^2), \\ \sigma_i^2 &\sim \mathcal{IG}(a_{i,\sigma^2}, b_{i,\sigma^2}), \end{aligned} \tag{3.10}$$

where $\mathcal{IG}$ indicates the inverse gamma distribution with scale parameter a_{i,σ^2} and shape parameter b_{i,σ^2}, and $\psi_{i,\chi}$ and w_χ^2 are respectively mean and variance of the normal distributions, for $\chi \in \{\mu, \gamma, a, \lambda\}$. The log transformations are used to ensure that the parameters are defined in the correct domain. In Equation (3.10), the means of the normal distributions are allowed to vary from patient to patient since there may exist covariates that affect them via a linear regression as follows,

$$\psi_{i,\chi} = \mathbf{C}_{i,\chi} \boldsymbol{\alpha}_\chi, \tag{3.11}$$

where $\mathbf{C}_{i,\chi}$ is a patient-specific vector of covariates of dimension p_χ and $\boldsymbol{\alpha}_\chi$ is a vector of regressors. Finally, covariate information can also be added to the component of the model that is used to define $\pi_i(t)$ by adding, in Equation (3.8), the following random intercept,

$$\beta_{0,i} = \mathbf{C}_{i,\beta}\boldsymbol{\alpha}_\beta. \tag{3.12}$$

The random effects $(a_{i,\sigma^2}, b_{i,\sigma^2}, \psi_{i,\chi}, w^2_{i,\chi})$, for $\chi \in \{\mu, \gamma, a, \lambda\}$, as well as the hyperparameters $\boldsymbol{\alpha}_\chi$, $\chi \in \{\mu, \gamma, a, \lambda, \beta\}$ are all appended to the parameter vector $\boldsymbol{\theta}$.

3.4 Estimating the optimal time

Equation (3.8) links the probability of a positive PET-PSMA test and the latent PSA level; this relation between the two is what we use to find the optimal time. We recall that the PSA level (when observed) is not the true underlying level, but a noisy version of the true latent one that cannot be directly measured.

3.4.1 Defining and identifying the optimal time

A solution to find the optimal time could be to plug in point estimates of all model parameters (maximum likelihood estimates, or Bayesian posterior means), and then invert Equation (3.8) to target the desired probability. This strategy is viable, but it does not take into account all sources of uncertainty. The Bayesian approach, which we follow globally to fit the model, gives a better way to estimate the optimal time and, simultaneously, account for uncertainty quantification in a controlled way.

In the Bayesian approach, the overarching goal is to compute the posterior density

$$f(\log \mathbf{y}^u, \mathbf{z}^u, \log \mathbf{x}^u, \log \mathbf{x}^o, \boldsymbol{\pi}^u, \boldsymbol{\pi}^o, \boldsymbol{\theta} \mid \log \mathbf{y}^o, \mathbf{z}^o), \tag{3.13}$$

based on which all quantities of interest may be computed. Marginalizing the posterior density in Equation (3.13), one can obtain, for each fixed time point t, the posterior predictive density

$$f\big(\pi_i(t), \tau_i \mid \log \mathbf{y}^o, \mathbf{z}^o\big). \tag{3.14}$$

It should be noted that this can be done since each specific τ_i is a component of the high dimensional parameter vector θ and, similarly, for each t, $\pi_i(t)$ is a parametric function. For each t, we can then verify whether the following condition is satisfied:

$$\mathcal{P}\big(\pi_i(t) > \pi^* \cap t > \tau \mid \log \mathbf{y}^o, \mathbf{z}^o\big) = \int_{\pi^*}^{1} \int_0^t f\big(\pi_i(t), \tau_i \mid \log \mathbf{y}^o, \mathbf{z}^o\big) d\tau_i d\pi_i(t) \geq \rho, \tag{3.15}$$

where:

- $\mathcal{P}$ stand for posterior predictive probability based on the density in Equation (3.14);
- π^* is a target probability of positive PET-PSMA;
- ρ is a posterior assurance probability (say 95%), similar to a confidence coefficient.

It should be noted that in Equation (3.15) we require t to be greater than the change point τ_i, whereas π^* and ρ are design parameters. The resulting decisions are

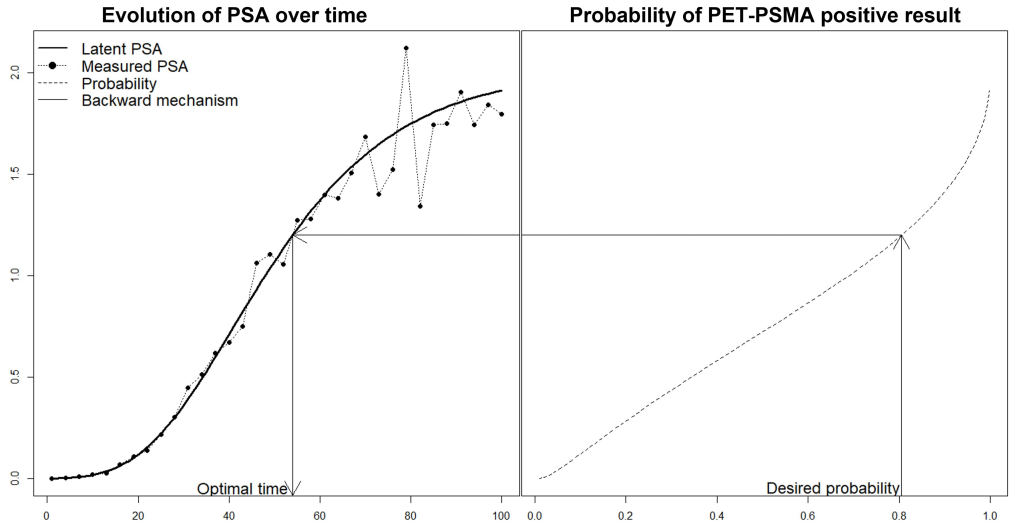

FIGURE 3.2: Joint modeling of $x_i(t)$ and $\pi_i(t)$. The plot shows the relation between time and PSA evolution and between PSA and the probability of a positive test. After choosing the desired probability, the associated PSA and time can be recovered through the model, following the arrows.

strongly dependent on the latter quantities, which should be carefully chosen in advance, in accordance with clinicians. Default values is $\rho = 0.95$, while π^* is usually selected with additional ROC analysis or according to clinicians expertise. Finally, to select the optimal time t_i^* for the i-th patient, we can choose the first available time in the future satisfying Equation (3.15), i.e., the smallest t greater than the largest time in $\mathcal{T}_i$ satisfying Equation (3.15). Figure 3.2 contains a graphical depiction of the procedure used to obtain the optimal time and how the latent PSA level is connected to the probability of a positive test. In particular, the plot shows the relation between time and PSA evolution, on the left side, and between PSA and the probability of a positive test, on the right side. After choosing the desired probability (on the right-side x-axis), the associated PSA and time can be recovered through the model, following the black arrows, from left to right.

3.4.2 Computing the optimal time

Due to the complexity of the hierarchical joint model, the posterior distribution is defined on high-dimensional data, which prevents us to compute, in closed form, any of the quantities that we may need for inference, as well as normalization constants. As often done with Bayesian models, we use Markov Chain Monte Carlo (MCMC) [Bro98] algorithms to obtain samples from the density in Equation (3.13) and Monte Carlo integration [GL06] approaches to approximate the posterior quantities of interest, such as posterior expectations and posterior probabilities and, specifically, the integral in Equation (3.15), which is the main focus of inference.

Let $\log \mathbf{y}^{u,b}, \mathbf{z}^{u,b}, \log \mathbf{x}^{u,b}, \log \mathbf{x}^{o,b}, \boldsymbol{\pi}^{u,b}, \boldsymbol{\pi}^{o,b}, \boldsymbol{\theta}^b$ be the b-th posterior samples from the associated parameters, where $b = 1, 2, \ldots, B$ and B is a large number of MCMC iterations. Using these posterior samples we can approximate the optimal-time t_i^*,

since samples from the predictive distribution in Equation (3.14) are easily obtainable. From Equations (3.5), (3.6) and (3.8) we can see that $\pi_i(t)$, for all t, is a deterministic function of the parameters, even if t is not in $\mathcal{T}_i$. Therefore, for a given t, and each b sample, we can compute

$$\log x_i^b(t) = \begin{cases} \lambda_i^b - \mu_i^b t, & \text{if } t < \tau_i^b, \\ \log x_i^b(\tau_i^b)\exp(-\gamma_i^b(t-\tau_i^b)) + a_i(1-\exp(-\gamma_i^b(t-\tau_i^b))), & \text{otherwise,} \end{cases} \tag{3.16}$$

and

$$\pi_i^b(t) = \frac{\left(x_i^b(t)\right)^{\beta_1^b} e^{\beta_{0,i}^b + \beta_2^b t}}{1 + \left(x_i^b(t)\right)^{\beta_1^b} e^{\beta_{0,i}^b + \beta_2^b t}}. \tag{3.17}$$

As a consequence, the set of samples $\{\pi_i^b(t), \tau_i^b\}_{b=1}^B$ are from the predictive distribution of Equation (3.14). It is easy to see that the integral in Equation (3.15) can be seen as an expectation if we rewrite it as

$$\begin{aligned} &\int_{\pi^*}^1 \int_0^t f\left(\pi_i(t), \tau_i \mid \log \mathbf{y}^o, \mathbf{z}^o\right) d\tau_i d\pi_i(t) = \\ &\qquad \int_0^1 \int_0^\infty 1_{[\pi^*,1]}\left(\pi_i(t)\right) 1_{[0,t)}(\tau_i) f\left(\pi(t), \tau_i \mid \log \mathbf{y}^o, \mathbf{z}^o\right) d\tau_i d\pi_i(t), \end{aligned} \tag{3.18}$$

where $1_{\cdot}(\cdot)$ is the characteristic function of a set. Hence, using samples from the predictive distribution, we can approximate the quantity in Equation (3.18) using a standard Monte Carlo integration, leading to the estimator

$$\begin{aligned} &\int_0^1 \int_0^\infty 1_{[\pi^*,1]}\left(\pi_i(t)\right) 1_{[0,t)}(\tau_i) f\left(\pi(t), \tau_i \mid \log \mathbf{y}^o, \mathbf{z}^o\right) d\tau_i d\pi_i(t) \approx \\ &\qquad \frac{\sum_{b=1}^B 1_{[\pi^*,1]}\left(\pi_i^b(t)\right) 1_{[0,t)}(\tau_i^b)}{B}. \end{aligned} \tag{3.19}$$

This means that the integral can be approximated by the proportion of posterior samples that are in the set $[\pi_i(t) > \pi^*, \tau_i < t]$. The approximation in Equation (3.19) must be computed for a fine grid of time points and the smallest $t \geq \max(\mathcal{T}_i)$ that satisfies

$$\frac{\sum_{b=1}^B 1_{[\pi^*,1]}\left(\pi_i^b(t)\right) 1_{[0,t)}(\tau_i^b)}{B} \geq \rho \tag{3.20}$$

is the desired optimal time.

3.5 Simulations

To conduct an investigation of the model and its practical estimation, and to emphasize the efficacy of the employed algorithm, we conduct a simulation study. Due to the complexity of the model, it may be useful to first investigate a single simulation and comment on its results, before going into the details of the full simulation study.

In this section, with a simulated dataset, we want to show that the model, and especially the change point values τ_i, can be estimated using the MCMC algorithm. We simulate a scenario where $m = 80$ patients are undergoing surgery at time 0 and then followed up for several months. For each patient, we simulate the number of elements of $\mathcal{T}_{y,i}$ from the distribution $U_d(5, 8)$, where $U_d(\cdot, \cdot)$ indicates a uniform

distribution over the integers between the two arguments (included). The times associated with the measurements are sampled randomly without replacement from the set $\{1, 2, \ldots, 25\}$, while the numbers of time points of the PET-PSMA exams are from the $U_d(3, 5)$ and the times are randomly sampled without replacement from the set $\{26, 27, \ldots, 38\}$. For each i, we assume $\min\{\mathcal{T}_{z,i}\} > \max\{\mathcal{T}_{y,i}\}$, meaning that the exams are always performed after the last PSA measurement and we have a small set of measurements for each patient. Let $\mathcal{T}_{y,i} = \{t_{i,j}\}_{j=1}^{n_i}$ be the set of ordered points: given these values, to simulate τ_i we sample from $U(t_{i,3}, t_{i,(n_i-2)})$, so that τ_i is in the middle of the temporal window. We simulate data in this way to create a challenging situation, where the data points are very few, and the PET-PSMA exams $\{z_i(t)\}_i$ are all performed much later than the last PSA measurement $\{y_i(t)\}_i$. This is done to highlight better how the method can be used to predict the PET-PSMA results for future time t, for which we have not observed the PSA level, and when data information is poor. For each patient i, we simulate 9 dichotomous variables $\{C_{ij}\}_{j=1}^{9}$, and a final variable $C_{i10} = \lfloor \hat{C}_{i10} \rfloor$, with $\hat{C}_{i10} \overset{i.i.d.}{\sim} \mathcal{N}(75, 7)$, used to describe the patient age. We then assume $\mathbf{C}_{i,\boldsymbol{\mu}} = \mathbf{C}_{i,\boldsymbol{\mu}} = (1, C_{i1}, C_{i2}, C_{i3}, C_{i4}, C_{i5})$ and $C_{i,\boldsymbol{\beta}} = (1, C_{i6}, C_{i7}, C_{i8}, C_{i9}, C_{i10})$. The remaining parameters are

$$\boldsymbol{\alpha_\mu} = \begin{pmatrix} 1 \\ 0.1 \\ 0.3 \\ 0.5 \\ 0.2 \\ 0.1 \end{pmatrix}, \quad \boldsymbol{\alpha_\gamma} = \begin{pmatrix} -1 \\ -0.01 \\ -0.01 \\ -0.01 \\ -0.01 \\ -0.01 \end{pmatrix}, \quad \boldsymbol{\alpha_\beta} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}, \quad \begin{pmatrix} \beta_1 \\ \beta_2 \\ \psi_a \\ \omega_{\boldsymbol{\mu}} \\ \omega_{\boldsymbol{\gamma}} \\ \omega_a \\ a_{\sigma^2} \\ b_{\sigma^2} \end{pmatrix} = \begin{pmatrix} 4 \\ 0.5 \\ 5.7 \\ 0.1 \\ 0.1 \\ 1 \\ 3 \\ 5 \end{pmatrix}. \tag{3.21}$$

Note that most of the parameters chosen for the simulation are randomly selected and there is no intention to be realistic. A Gaussian prior $\mathcal{N}(0, 100)$ is used on

$$\lambda_i,\ \log \omega_\mu,\ \log \omega_\gamma,\ \psi_a,\ \log \frac{b_{\sigma^2}}{(a_{\sigma^2} - 1)},\ \log \frac{b_{\sigma^2}^2}{(a_{\sigma^2} - 1)^2(a_{\sigma^2} - 2)}, \tag{3.22}$$

and all regression coefficients, where $b_{\sigma^2}/(a_{\sigma^2} - 1)$ and $b_{\sigma^2}^2/\left((a_{\sigma^2} - 1)^2(a_{\sigma^2} - 2)\right)$ are respectively mean and variance of the parameter σ_i^2. The prior of τ_i is a patient-specific mixed-type distribution: τ_i can assume value in $\{t_{i1}, [t_{i2}, t_{iJ_{i-1}}], t_{iJ_i}\}$, where $\{t_{i1}, t_{i2}, ..., t_{iJ_i}\} \equiv \mathcal{T}_{y,i}$, with a probability mass of 1/3 on t_{i1} and t_{iJ_i}, and a Uniform distribution in $[t_{i2}, t_{iJ_{i-1}}]$. The reasoning behind this choice is that a change point in $[0, t_{i2})$ is not identifiable, since only the observation collected at time t_{i1} is available to estimate the decreasing phase. The same applies in the time window $(t_{iJ_{i-1}}, +\infty)$ where we could gain information only from t_{iJ_i} to estimate the log Gompertz coefficients. Since the change point is one of the most important parameters for the final identification of the optimal time, we prefer to have a parameter that is always identifiable. We run our algorithm for 150.000 iterations, with burn-in equal to 100.000 and thinning parameter equal to 10. The algorithm is a Metropolis-within-Gibbs MCMC with the adaptive Metropolis steps proposed in [AT08], and a Pólya-Gamma update [PSW13] for $(\beta_1, \beta_2, \alpha_\beta)$.

We sum up the results in two tables. In Table 3.1, for each individual parameter, we compute the posterior 95% credible interval (CI) and we evaluate the percentage of parameters correctly estimated (for which the true value used for simulation

falls in the 95% CI), across individuals. On the other hand, Table 3.2 sums up the results obtained for the global parameters, in particular, it reports the true values, the posterior mean, and the 95% CIs. From Table 3.1, we see good performances on individual parameters' estimation, as expected the performance on change points τ_i is more challenging, but the results are still accurate, compared with other individual parameters. From Table 3.2 we can see that 24 out of 26 global parameters are well estimated as the true value is contained in the 95% CI.

3.5.1 Full simulation study

The full simulation study comprises four scenarios: scenario 1 and scenario 2 differ only in the noise value (determined by parameters a_{σ^2} and b_{σ^2}), while the third scenario differs in most of the parameters. Scenario 4 closely resembles scenario 1 but is composed of fewer data points. Specifically, in scenarios 1-2-3, the datasets consist of 80 patients, with each patient having between 8 and 15 PSA measurements and between 3 and 5 PET-PSMA examinations. In scenario 4, the dataset is composed of 80 patients, with each patient having between 6 and 10 PSA measurements, and between 1 and 3 PET-PSMA examinations. The underlying parameters used to simulate the datasets are summarized in Table 3.3. For each of the four scenarios, we simulate 100 datasets and run the MCMC procedure to obtain samples from the posterior. The results are summarized in Tables 3.4 and 3.5, reporting the percentage of global and individual parameters correctly estimated (the true values are within the posterior 95% interval). Additionally, the tables include information about these intervals lengths (at levels 0.25, 0.5, and 0.975). Overall, the performances are satisfactory. Moreover, the simulation study highlights that higher values of the noise term lead to a decrease in the model ability to estimate the growth coefficients μ_i and γ_i, which can be seen by interval lengths in Table 3.5. However the general performance of the model is still satisfactory. The simulation study, in particular, underscores how the quality of estimation and the result reliability strongly depend on the number of available data points for each patient: if the number of collected PET-PSMA measurements decreases, although the percentage of correctly identified parameters remains relatively high, the posterior credible interval amplitude increases, resulting in reduced precision in estimating the parameters for the logistic model.

TABLE 3.1: Simulated dataset - Results on individual parameters: percentage of correctly identified individual parameters out of 80 patients.

Parameter	Percentage
$\log(\lambda_i)$	95.00%
τ_i	95.00%
$\log(\mu_i)$	91.25%
$\log(\gamma_i)$	95.00%
$\log(a_i)$	100%
σ_i^2	96.25%

TABLE 3.2: Simulated dataset - Results global parameters.

Parameter	2.5%	Mean	97.5%	Real
$\alpha_\mu[1]$	0.903	0.978	1.055	1.00
$\alpha_\mu[2]$	0.012	0.075	0.134	0.10
$\alpha_\mu[3]$	0.233	0.296	0.355	0.30
$\alpha_\mu[4]$	0.459	0.519	0.578	0.50
$\alpha_\mu[5]$	0.131	0.188	0.247	0.20
$\alpha_\mu[6]$	0.091	0.149	0.213	0.10
$\alpha_\gamma[1]$	-1.074	-0.952	-0.844	-1.00
$\alpha_\gamma[2]$	-0.119	-0.039	0.042	-0.01
$\alpha_\gamma[3]$	-0.045	0.025	0.098	-0.01
$\alpha_\gamma[4]$	-0.034	0.047	0.171	-0.01
$\alpha_\gamma[5]$	-0.076	-0.006	0.079	-0.01
$\alpha_\gamma[6]$	-0.151	-0.072	0.003	-0.01
$\beta[1]$	2.486	4.809	8.742	4.000
$\beta[2]$	0.500	0.859	1.346	0.500
$\alpha_\beta[1]$	-6.384	7.025	20.764	1.00
$\alpha_\beta[2]$	-0.482	2.492	6.046	1.00
$\alpha_\beta[3]$	-1.820	1.121	4.564	1.00
$\alpha_\beta[4]$	-2.099	0.878	4.276	0.50
$\alpha_\beta[5]$	-4.558	-1.242	1.597	-0.50
$\alpha_\beta[6]$	-1.213	-0.783	-0.463	-0.50
ω_μ	0.084	0.109	0.136	0.10
ω_γ	0.059	0.103	0.157	0.10
ψ_a	4.825	5.582	6.405	5.70
ω_a	0.837	1.272	1.915	1.00
a_{σ^2}	2.000	2.239	9.486	3.00
b_{σ^2}	3.671	5.276	9.486	5.00

TABLE 3.3: Four simulated scenarios: global parameters of interest used to simulate the dataset.

	α_μ						α_γ					
Scen. 1	0.5	0.1	0.3	0.5	0.2	0.1	-0.5	-0.1	-0.1	-0.1	-0.1	-0.1
Scen. 2	0.5	0.1	0.3	0.5	0.2	0.1	-0.5	-0.1	-0.1	-0.1	-0.1	-0.1
Scen. 3	0.1	0.3	0.3	0.5	0.2	0.5	-0.1	-0.5	-0.5	-0.5	-0.5	-0.5
Scen. 4	0.5	0.1	0.3	0.5	0.2	0.1	-0.5	-0.1	-0.1	-0.1	-0.1	-0.1
	α_β						β	ψ_a	ω_a	$\omega_{\mu,\gamma}$	a_{σ^2}	b_{σ^2}
Scen. 1	0.5	1	1	0.5	-0.5	-0.5	(1; 2)	2.48	1	0.1	3	5
Scen. 2	0.5	1	1	0.5	-0.5	-0.5	(1; 2)	2.48	1	0.1	3	15
Scen. 3	0.5	0.5	0.5	0.1	-0.1	-0.1	(0.2; 4)	2.48	1	0.5	2	7
Scen. 4	0.5	1	1	0.5	-0.5	-0.5	(1; 2)	2.48	1	0.1	3	5

TABLE 3.4: Four simulated scenarios, for each 100 dataset are analyzed. Table reports the percentage of global parameters correctly contained in their posterior estimation intervals, in addition the 95% quantile with the median value for the intervals length is reported in square brackets.

Par.	Scenario 1	Scenario 2	Scenario 3	Scenario 4
$\alpha_\mu[1]$	91%	87%	90%	91%
	[0.13 0.17 0.22]	[0.15 0.21 0.29]	[0.10 0.30 0.30]	[0.13 0.17 0.22]
$\alpha_\mu[2]$	95%	95%	94%	95%
	[0.10 0.13 0.17]	[0.11 0.15 0.20]	[0.52 0.66 0.82]	[0.10 0.13 0.17]
$\alpha_\mu[3]$	97%	94%	98%	97%
	[0.10 0.13 0.16]	[0.11 0.16 0.21]	[0.45 0.51 0.60]	[0.10 0.13 0.16]
$\alpha_\mu[4]$	97%	97%	93%	97%
	[0.10 0.13 0.16]	[0.12 0.15 0.23]	[0.43 0.52 0.60]	[0.10 0.13 0.16]
$\alpha_\mu[5]$	96%	94 %	94%	96%
	[0.10 0.13 0.17]	[0.11 0.15 0.20]	[0.44 0.52 0.59]	[0.10 0.13 0.17]
$\alpha_\mu[6]$	95%	95%	94 %	95%
	[0.10 0.13 0.16]	[0.11 0.16 0.23]	[0.44 0.51 0.60]	[0.10 0.13 0.16]
$\alpha_\gamma[1]$	100%	99%	97 %	100 %
	[0.30 0.39 0.61]	[0.28 0.48 0.76]	[0.65 0.80 0.99]	[0.30 0.39 0.61]
$\alpha_\gamma[2]$	99%	98%	97 %	99%
	[0.19 0.24 0.36]	[0.16 0.27 0.39]	[0.45 0.57 0.67]	[0.19 0.24 0.36]
$\alpha_\gamma[3]$	97%	91%	94 %	97%
	[0.19, 0.24 , 0.33]	[0.14 0.29 0.44]	[0.45 0.58 0.69]	[0.19 0.24 0.33]
$\alpha_\gamma[4]$	98%	96%	95 %	98%
	[0.19 0.25 0.37]	[0.17 0.30 0.51]	[0.47 0.58 0.69]	[0.19 0.25 0.37]
$\alpha_\gamma[5]$	99%	92%	96 %	99%
	[0.19 0.25 0.31]	[0.14 0.29 0.43]	[0.47 0.57 0.72]	[0.19 0.25 0.31]
$\alpha_\gamma[6]$	99%	97%	96 %	99%
	[0.19 0.25 0.34]	[0.15 0.28 0.44]	[0.48 0.58 0.70]	[0.19 0.25 0.34]
$\beta[1]$	60%	61%	26 %	60%
	[4.80 13.95 2426]	[11.73 25.28 31.50]	[11.73 25.28 31.50]	[4.80 13.95 24.45]
$\beta[2]$	94%	95%	89 %	55 %
	[0.05 0.07 0.13]	[0.07 0.11 0.27]	[3.34 7.38 11.52]	[0.54 1.49 9.31]
$\alpha_\beta[1]$	100%	100%	100%	100%
	[25.18 36.32 39.44]	[26.06 36.20 39.26]	[38.23 39.04 40.10]	[25.18 36.32 39.44]
$\alpha_\beta[2]$	95 %	94%	100 %	95%
	[4.54 13.73 30.38]	[4.58 13.86 31.50]	[23.07 31.95 37.77]	[4.54 13.73 30.55]
$\alpha_\beta[3]$	92%	92%	100%	92%
	[4.34 14.12 30.32]	[4.48 14.58 30.72]	[21.06 32.27 35.65]	[4.34 14.12 31.00]
$\alpha_\beta[4]$	94%	93%	100%	94%
	[4.42 13.22 30.24]	[4.57 13.62 30.62]	[23.31 32.14 36.10]	[4.42 13.22 30.40]
$\alpha_\beta[5]$	98%	98%	100%	98%
	[4.16 13.34 30.44]	[4.30 13.73 30.49]	[21.71 32.04 36.27]	[4.16 13.34 30.94]
$\alpha_\beta[6]$	48%	50%	95 %	48%
	[0.60 1.82 6.62]	[0.64 1.81 6.81]	[1.39 2.99 4.52]	[0.60 1.82 6.62]
ω_μ	56%	82%	99 %	94%
	[0.11 0.15 0.27]	[0.15 0.22 0.39]	[0.18 0.21 0.26]	[0.05 0.07 0.13]

Continued on next page

Table 3.4 – Continued from previous page

Par.	Scenario 1	Scenario 2	Scenario 3	Scenario 4
ω_γ	56%	54%	99 %	57%
	[0.54 1.46 9.31]	[0.55 1.49 11.67]	[0.20 0.26 0.34]	[0.11 0.15 0.27]
ψ_a	98%	100%	99%	98%
	[1.19 1.74 3.05]	[1.45 2.21 4.66]	[1.11 1.69 5.30]	[1.19 1.74 3.05]
ω_a	99%	100%	95 %	99%
	[0.70 1.10 1.94]	[0.86 1.29 2.40]	[1.08 3.28 10.28]	[0.70 1.10 1.94]
a_{σ^2}	95%	97%	100 %	95%
	[0.75 2.55 7.12]	[0.86 2.92 8.37]	[0.19 0.61 2.26]	[0.75 2.54 7.12]
b_{σ^2}	100%	98%	78%	100%
	[3.25 7.71 19.09]	[9.92 22.83 61.88]	[3.51 5.26 13.38]	[3.25 7.58 19.09]

TABLE 3.5: Four simulated scenarios, for each 100 dataset are analyzed. Table reports the percentage of individual parameters correctly contained in their posterior estimation intervals, in addition the 95% quantile with the medial value for the intervals length is reported in square brackets.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4
$\log(\lambda_i)$	95%	93%	94%	95%
	[3.11 5.17 9.96]	[4.38 7.04 12.79]	[4.11 8.08 18.63]	[3.11 5.17 9.98]
τ_i	98%	96%	95%	98%
	[0.26 0.90 3.52]	[0.40 1.25 4.58]	[0.34 1.78 6.99]	[0.26 0.90 3.52]
$\log(\mu_i)$	98%	90%	96%	98%
	[0.10 0.25 0.44]	[0.00 0.27 0.52]	[0.01 0.22 1.93]	[0.10 0.25 0.43]
$\log(\gamma_i)$	95%	90%	93%	95%
	[0.30 0.67 1.45]	[0.33 0.79 1.67]	[0.40 1.16 3.90]	[0.30 0.67 1.45]
$\log(a_i)$	94%	93%	97%	94%
	[2.40 3.90 6.34]	[2.72 4.42 8.04]	[2.56 4.39 7.10]	[2.40 3.90 6.35]
σ_i^2	88%	91%	94%	88%
	[0.94 1.53 2.88]	[1.39 2.08 3.98]	[1.07 1.77 4.04]	[0.94 1.53 2.89]

3.6 Application to clinical data

The database is built on clinical practice in the San Luigi Hospital in Torino. In prostate cancer follow-up setting post radical prostatectomy (RP), PSA is the main source of information to base clinical decisions regarding interventions such as PET-PSMA exam. We have $n = 111$ patients and, for each of them, we have several demographic and clinical variables, as well as the PSA measurements and PET-PSMA results taken at different times after RP. The number of time points in $\mathcal{T}_{y,i}$ ranges from 4 to 17, while the ones in $\mathcal{T}_{z,i}$ from 1 to 4, and the follow-up period is between 4 and 280 months. We can distinguish between patients who, after prostatectomy, show relatively low values of PSA after RP, but may be eventually subject to a biochemical relapse at change point (BCR patients), and Biochemical-persistence patients (BCP patients), who present persistent benign/malignant residual tissue after surgery (conventionally signaled by PSA > 0.2) and are usually treated with therapy to inhibit cancer growth - for them, PSA levels may even initially decrease until a

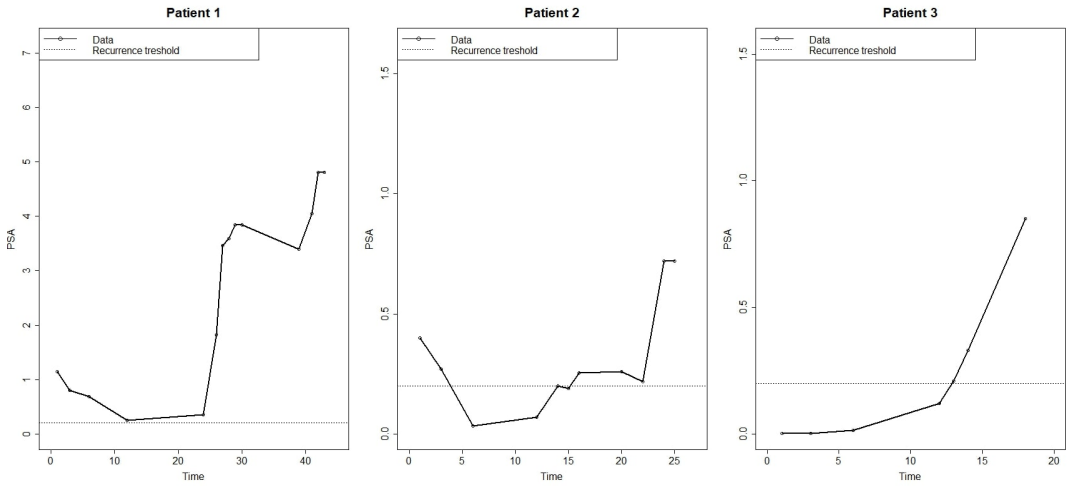

FIGURE 3.3: PSA measurements collected on three patients. Patients 1 and 2 are BCP but with quite different PSA levels, while 3 is BCR.

new increase occurs at the change point. The presence of a mixed population of BCR and BCP patients and the possible administration of different therapies make the inference task a very difficult exercise. To present the possible data paths, Figure 3.3 shows data collected for some anonymous patients, that we refer to from now on as patients 1, 2, and 3. The shape of the data for these patients is quite different (patients 1 and 2 are BCP, while patient 3 is BCR), but also similar shapes can correspond to quite different PSA values (patient 1 and patient 2). Before surgery, each patient is assigned to a different category according to the status (S_i), but some more information is collected at surgery time; in particular, pathological stadiation according to TNM classification, and the clinical status of prostate margins. Regarding tumor status P_i^T, the clinicians use four different categories: $T1$ (clinically unapparent tumor), $T2$ (tumor confined within prostate), $T3$ (tumor extends through the prostate capsule), $T4$ (tumor is fixed or invades adjacent structures). For the sake of simplicity, here the 4 tumor categories have been reduced to two: $P_i^T = 0$ if T_1 and T_2 and $P_i^T = 1$ otherwise. The lymph nodes implication is evaluated through a binary variable (P_i^N), while metastases are always present and, thus, not relevant. Moreover, the clinical stage of the tumor is usually represented using the Gleason Score [Ege+02], and we reduced the original nine categories to two: $S_i = 0$ if Gleason Score is less than 6 or if it is equal to 3+4, $S_i = 1$ otherwise. Finally, prostate resection margins (P_i^R) are evaluated after the surgery.

According to clinical evaluation, each patient with BCR and/or BCP can then be administered four different therapies: adjuvant or salvage androgen deprivation hormone therapy (OA_i or OS_i), adjuvant or salvage radiotherapy (RA_i or RS_i). Some patients also underwent regional lymphadenectomy (L_i). Adjuvant therapies are administered within 6 months from surgery, while salvage ones are performed after six months on from surgery. In addition, for each patient, we know the age A_i. All these data are used as clinical and demographic covariates to gain information useful for our model. In particular, we assume that μ_i, γ_i, σ_i^2 and a_i are random effects, with $\psi_{i,a} = \psi_a$, a_{i,σ^2}=a_{σ^2}, and b_{i,σ^2}=b_{σ^2} (constant for each patient), while the means of both μ_i, γ_i depend on covariates, to gain more information from the available data. However, the observed difference between the reported categories of patients, namely BCR and BCP patients, suggests that the parameters λ_i do not come from a common

population. Instead of constructing a bimodal random effect which is out of our scope, we take each λ_i to be an individual parameter. We estimated three models with different combinations of covariates $\mathbf{C}_\mu$, $\mathbf{C}_\gamma$, $\mathbf{C}_\pi$, all suggested by clinical evidence, and among them, we selected the best model using the Watanabe–Akaike information criterion (WAIC) [GHV14], see Table 3.6. The results we describe refer to the best configuration selected, namely Model 2. In Tables 3.7, 3.8, 3.9, and 3.10 we show the global parameter estimates for the best model, with relative 95% CIs. Interpretation of coefficients can be difficult and misleading, particularly when referring to therapies. These are the usual difficulties in interpreting causality with observational and clinical practice databases rather than clinical trials. From a clinical perspective, one could expect therapies to decrease the PSA level before change point τ and to reduce the growth speed after it. However, the analysis is not targeting therapeutic efficacy. Therapies are used here as indicators of the severity of the disease rather than for estimating their effect. For this reason, we report the logistic analyses used to determine the relationship between administrations of therapies and baseline covariates (see Table 3.11), where it is easy to notice that, as expected, patients with the worst clinical situation at the surgery time have higher probabilities of receiving one or more of the analyzed therapies. On the other hand, baseline covariates have relatively simple interpretations. Following the theory, PSA decreasing level turns out to be almost zero for patients that do not receive therapies (see $\alpha_\mu[1]$). Gleason score is a significant factor, as higher levels determine an increase in PSA values. In particular, the standard deviation of decreasing coefficient ω_μ and increasing coefficient ω_γ are quite different. The high level of the former reflects the heterogeneity of the dataset: PSA levels right after surgery and preceding the changing points are different for BCR and BCP. Finally, the probability of a positive PET-PSMA result at time t strongly depends on the PSA level at t (as was to be expected), on the Gleason score, and on lymph nodes implication. In particular, as the PSA level increases, also the probability increases.

TABLE 3.6: Different configuration of covariates. □ indicates values included in the mean of μ; △ indicates values included in the mean of γ; × indicates values included in the logistic regression, namely π; for each configuration the WAIC is reported.

Model	OA	RA	OS	RS	L	P^R	P^T	P^N	S	A	WAIC
1	□△×	□△×	□△×	□△×	□△×	□△×	□△×	□△×	□△×	□△×	1230819
2	□△	□△	□△	□△	□△	△×	△×	△×	△×	△×	1093648
3	□	□	△	△	□△	△×	△×	△×	△×	△×	1205765

TABLE 3.7: Clinical dataset - Results on global parameters for PSA growth curve before change point. Positive coefficients are associated with a lower level of PSA.

Parameter	2.5%	Mean	97.5%
$\alpha_\mu[1]$ - Intercept	-10.37	-7.76	-5.81
$\alpha_\mu[2]$ - Ormono adj	-0.20	1.84	3.84
$\alpha_\mu[3]$ - Ormono salvage	-5.67	-3.18	-0.49
$\alpha_\mu[4]$ - Radio adj	-1.66	0.58	2.85
$\alpha_\mu[5]$ - Radio salvage	-0.53	1.46	3.36
$\alpha_\mu[6]$ - Lymphadenectomy	-1.30	1.13	4.25

TABLE 3.8: Clinical dataset - Results on global parameters for PSA growth curve after the change point. Positive coefficients are associated with a higher level of PSA.

Parameter	2.5%	Mean	97.5%
$\alpha_\gamma[1]$ - intercept	-2.99	-2.58	-2.18
$\alpha_\gamma[2] - P^R$	-0.14	0.20	0.56
$\alpha_\gamma[3] - P^T$	-0.29	0.08	0.44
$\alpha_\gamma[4] - P^N$	-0.78	-0.17	0.47
$\alpha_\gamma[5] - S$	0.32	0.64	1.03
$\alpha_\gamma[6]$ - Age	-0.25	-0.04	0.14
$\alpha_\gamma[7]$ - Ormono adj	0.18	0.71	1.28
$\alpha_\gamma[8]$ - Ormono slv	0.26	0.73	1.28
$\alpha_\gamma[9]$ - Radio adj	-1.69	-1.13	0.547
$\alpha_\gamma[10]$ - Radio Slv	-0.65	-0.24	0.21
$\alpha_\gamma[11]$ - Lymphadenectomy	-0.41	0.03	0.49

TABLE 3.9: Clinical dataset - Results on global parameters for logistic model. Positive coefficients are associated with higher probabilities of positive PET-PSMA examinations.

Parameter	2.5%	Mean	97.5%
$\alpha_\beta[1]$ - Intercept	-0.59	0.84	2.34
$\alpha_\beta[2] - P^R$	-0.11	0.68	1.62
$\alpha_\beta[3] - P^T$	-0.84	0.22	0.85
$\alpha_\beta[4] - P^N$	0.05	0.88	2.56
$\alpha_\beta[5] - S$	0.05	1.28	2.56
$\alpha_\beta[6]$ - Age	-0.52	-0.14	0.24
β_1	1.64	2.56	3.68
β_2	-0.01	0.00	0.01

TABLE 3.10: Clinical dataset - Results on global parameters.

Parameter	2.5%	Mean	97.5%
ω_μ	2.53	3.14	3.91
ω_γ	0.63	0.75	0.90
Ψ_a	-0.41	-0.26	-0.13
ω_a	0.57	0.66	0.76
a_{σ^2}	2.00	2.01	2.02
b_{σ^2}	0.18	0.22	0.27

TABLE 3.11: Clinical dataset - Coefficients of logistic regression on probability to receive therapies. Bold coefficients refer to p-values smaller than 0.005.

Therapy	Intercept	P^R	P^T	P^N	S	L	PSA at surgery
Ormono adjuvant	**-3.22**	-0.02	-0.46	**1.92**	1.20	0.46	**1.54**
Ormono salvage	**-1.57**	-0.08	-0.95	-0.51	0.90	0.54	-1.39
Radio adjuvant	**-2.94**	**2.21**	-0.57	0.92	0.35	-0.30	-0.25
Radio salvage	**-3.14**	0.52	1.02	-0.70	1.01	0.23	-0.79
Lymphadenectomy	-0.03	-0.26	0.84	N.A.	**1.12**	**-0.60**	0.25

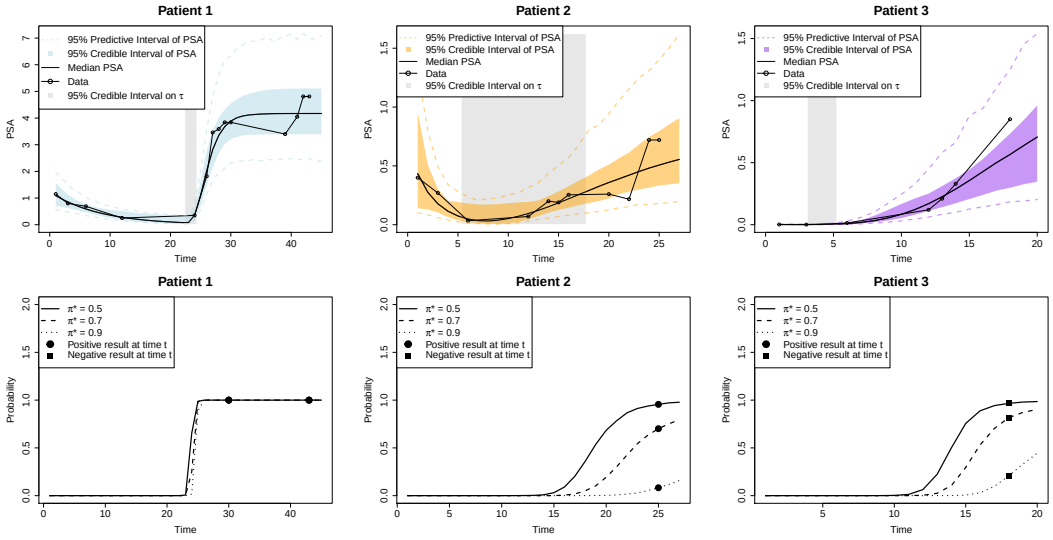

FIGURE 3.4: PSA growth curves and probability curves for the three patients introduced in Figure 3.3. Patients 1 and 2 are BCP but with quite different PSA levels, while 3 is BCR.

In Figure 3.4 we show examples of the model outputs. In the first row, the fitted curves for the selected patients (first row) are shown, where the grey solid regions are used to show the 95% CI if τ_i. In the second row, the approximation on the right-hand side of Equation (3.19) is computed using thresholds 0.5, 0.7, and 0.9: as expected, the higher the threshold, the lower the probability. Filled circles and squares are used to indicate the true and negative outcomes of the test, respectively. The posterior distributions of individual parameters λ_i, τ_i, γ_i, μ_i, a_i are reported in Figure 3.5. The posterior distributions λ_i show differences that reflect the difference between BCP and BCR patients. In particular, it is easy to see that we obtain different fits for patients 1 and 2 of type BCP, in comparison with BCR patients 3.

In the PET-PSMA case study presented, data are collected in an observational setup and not in a clinical trial. Therefore the time-grid of measurements (in particular on PET-PSMA) is not prespecified, but affected by the clinician decision and the severity of the illness itself, sources of potential bias. However, our work aims to develop a general tool to be used in several scenarios, not to construct a model to analyze the specific dataset presented, which served more as a case study. To test and promote our new method, the final idea is to construct a general database, that can be accessed by multiple clinicians, containing patients from multiple centers, whose information should be updated and added at each new visit, to enforce the power of the statistical methodology. Clearly, the data collected with this procedure are heterogeneous and we try to model them in a realistic and parsimonious way.

3.6.1 Comparison of the joint model to the logistic model

We compare the performance of our joint model to the closest method existing in the literature [Lui+20], that can be considered the current standard of care. In particular, a logistic regression model is fit with the `R` function `glm` [R C] (following the idea proposed in [Lui+20]), including the same baseline covariates as our joint-logistic model and the observed PSA level. In this logistic model, which we compare to our joint model, the response is 1 if the PET-PSMA gives a positive result (i.e., the

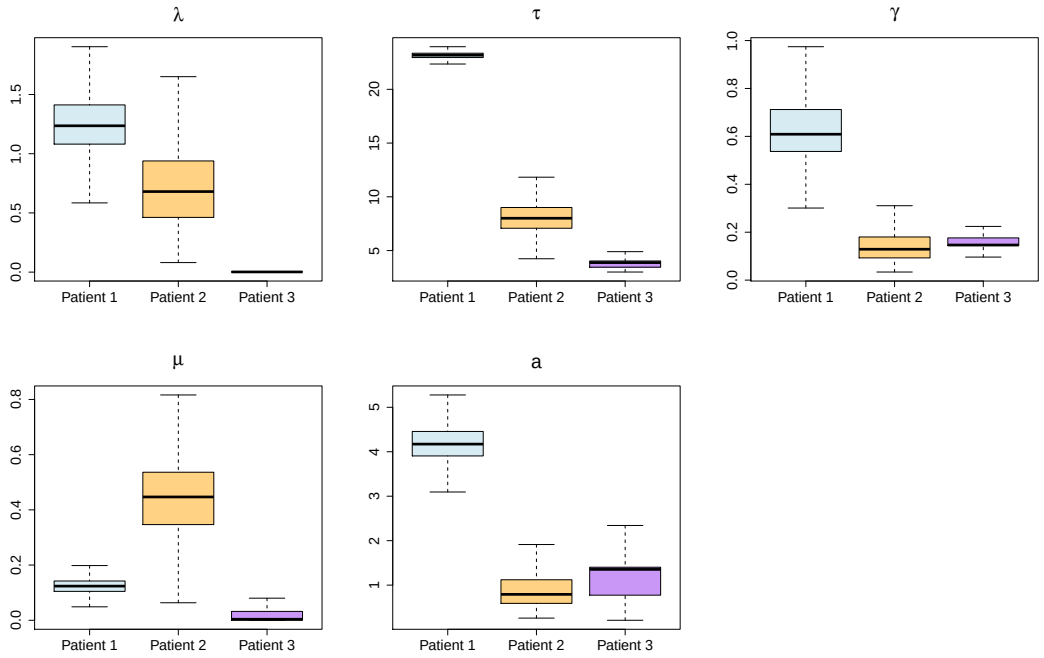

FIGURE 3.5: Individual parameters posterior distributions for the three patients introduced in Figure 3.3.

location of the disease is identified) and 0 if not. We apply to the San Luigi Hospital dataset presented in Section 3.6 both methods. Overall, the results obtained from this logistic analysis (see Table 3.12) are largely consistent with those from our method. Moreover, the significant covariates are also overall consistent with the reference paper [Lui+20], based on a different but similar dataset. Specifically, we found that tumor status and log-PSA level have strong and significant effects on the probability of positive PET-PSMA results, while resection margin and time have weaker effects. Notably, all of these covariates have positive coefficients, indicating that higher values increase the likelihood of positive PET-PSMA results. To formally compare the two fitted models, we use them to predict the PET-PSMA binary outputs, on the dataset under analysis, and we evaluate the relative Receiver Operating Characteristic curves (ROCs), shown in Figure 3.6, computing their areas under the curve (AUCs). Figure 3.6 shows the ROC curves for the logistic and the joint model,

TABLE 3.12: R output of the logistic model based on the idea of [Lui+20] applied to the medical data.

	Estimate	Std. Error	z value	Pr($>z$|)
(Intercept)	-0.2754	0.4757	-0.58	0.5627
P^R	0.6082	0.3622	1.68	0.0931
P^T	0.0918	0.3613	0.25	0.7994
P^N	0.3738	0.5504	0.68	0.4971
S	1.0768	0.3628	2.97	0.0030
Age	-0.1381	0.1649	-0.84	0.4023
log(PSA)	1.3855	0.2628	5.27	0.0000
time	0.0070	0.0042	1.69	0.0915

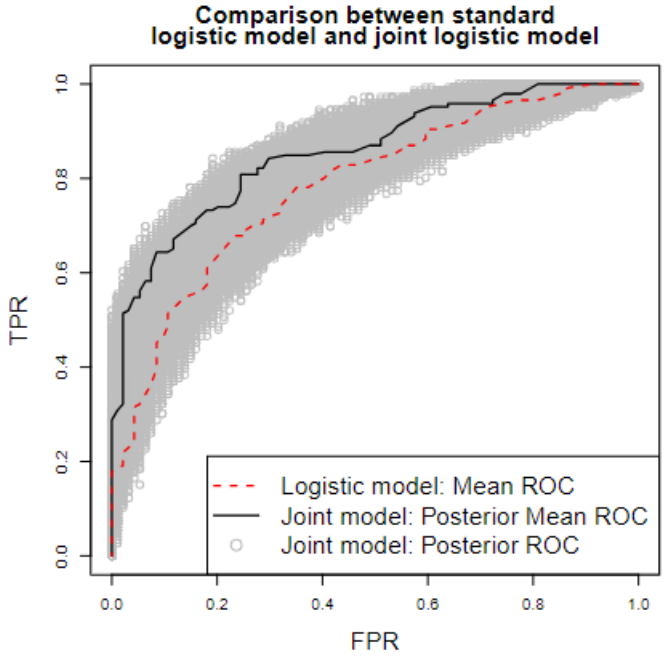

FIGURE 3.6: ROC curves for simple logistic model and join logistic model on real data. The AUC for simple logistic ROC is 0.79, the AUC for mean joint model ROC is 0.86, while the posterior 95% distribution of the AUC index ranges between 0.78 and 0.86. Results are obtained with `R` package `pROC` [Rob+].

computed with real data using `R` function `auc` of the `pROC` package [Rob+]. The AUC for the simple logistic (red line) is 0.79, while the AUC for mean joint model (black line) is 0.86, with the 95% of posterior AUC values ranging between 0.78 and 0.85, with median 0.81. As a result, we can see that the use of the latent PSA level, i.e., $x_i(t)$, one of the main points of our proposal, instead of the measured level, which is used in the competing model [Lui+20], gives better results. From the mean ROC curves, we can select the optimal π^* probability, namely the one corresponding to the higher trade-off between sensitivity and specificity; it is selected as the point on the ROC closest to coordinates $(0, 1)$, which corresponds to $\pi^* = 0.55$. Moreover, one big advantage of our joint model is that it enables the prediction of probabilities of positive results for future times, for which the PSA level may still be unknown. For example, this prediction of the PSA level and the related probability of positive success can still impact the treatment strategy adopted in the meantime, thus improving the patient benefit.

In addition to the logistic regression presented in [Lui+20] and analyzed in this Section, a common procedure used by clinicians is the evaluation of PSA-doubling time (PSA-DT) and PSA-velocity, based on the last couple of PSA measurements. In literature, PSA-DT higher than six months is significantly associated with a high probability of positive PET-PSMA results, but no clear and well-defined guidances are available, resulting in subjective decisions of time to examination, which depend on the clinician belief and past experiences. For this reason, a comparison of the performances of our proposed joint model and this methodology cannot be performed.

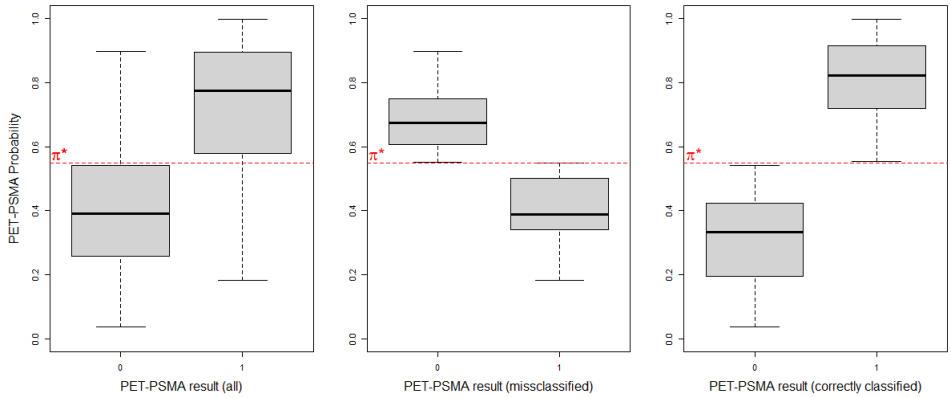

FIGURE 3.7: Predictive probability of positive PET-PSMA results for observed negative (0) and positive (1) examinations of the whole dataset (left panel), of misclassified results only (middle panel), and correctly classified results (right panel), obtained through cross-validation (with $\pi^* = 0.55$).

3.6.2 A Cross-Validation study

To evaluate the performance of the proposed algorithm on the real dataset, we additionally perform a leave-one-out cross-validation analysis, according to Section 1.2. We exploit the cross-validation predictive density sets

$$\{f(y_i(t)|\log \mathbf{x}^o, \boldsymbol{\pi}^o, \log \mathbf{y}^o_{it}, \mathbf{z}^o); \quad \forall\, y_i(t) \in \mathbf{y}^o\},$$

for the PSA levels, where $\log \mathbf{y}^o_{it}$ is the set of observed log-PSA values except for $\log y_i(t)$, and

$$\{f(z_i(t)|\log \mathbf{x}^o, \boldsymbol{\pi}^o, \log \mathbf{y}^o, \mathbf{z}^o_{it}); \quad \forall\, z_i(t) \in \mathbf{z}^o\},$$

for the PET-PSMA examination, where $\mathbf{z}^o_{it}$ is the set of observed log-PSA values except for $z_i(t)$. For each PSA and PET-PSMA measurement collected, we sample 1000 realizations from the corresponding cross-validation predictive density, estimated through 150.000 iterations of the algorithm (burn-in 100.000, thinning parameter 10). To evaluate the goodness of the prediction we use the accuracy index. The PSA accuracy, i.e. the percentage of predictive 95% CI containing the true values of the corresponding measurement, equals 98.20%. On the other side, PET-PSMA accuracy, i.e. the percentage of PET-PSMA results correctly predicted (threshold $\pi^* = 0.55$, as selected with ROC analysis in Section 3.6.1), equals 77.08%. Moreover, to validate the posterior distribution of $\pi_i(t)$, Figure 3.7 shows the probability of positive results across different subsets (0 for negative and 1 for positive) PET-PSMA results. In the left panel, we present the overall dataset, the middle panel focuses on misclassified instances, and, finally, the right panel highlights correctly classified cases. The plot shows that the posterior median of $\pi_i(t)$, respectively for the outcome 0 and 1, is approximately 0.4 and 0.78 when considering the entire dataset. However, when we restrict our analysis to the correctly characterized data, these values shift to ≈ 0.35 for 0 and ≈ 0.82 for 1, which demonstrates that the right decision can be made with a high degree of confidence. In the middle panel, when the method encounters challenges in classifying the data, there is a notable increase in uncertainty. The values of $\pi_i(t)$ are close to the threshold π^*, confirming that the methods fail when uncertainty is high. This underscores the sensitivity of the model to situations with increased ambiguity and suggests that the uncertainty in predictions increases

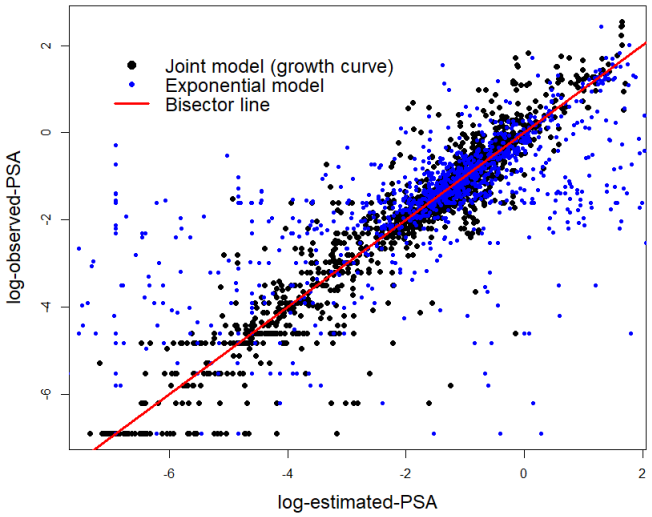

FIGURE 3.8: Mean estimated log-psa (on x-axis) compared with observed log-psa (y-axis): black points are estimated through our joint model, blue points are estimated through exponential model. Red line is the bisector. Results are obtained through cross-validation.

when the method encounters difficulty in making accurate classifications.

We compare the accuracy of our joint model with results obtained with standard methods, consistently with existing literature practices. We conduct leave-one-out cross-validation to assess the performance of the PET-PSMA logistic model (based on [Lui+20] and presented in Section 3.6.1), but also the exponential-fit model, a simpler method which predicts the future (and yet unknown) PSA level at time t_3 based on pairs of PSA measurements (PSA_1, PSA_2) collected at previous times t_1 and t_2, utilizing an exponential fit [Ben+08; MSK]. Note that the PSA level at the given time t_3, is predicted by only using two previous recorded values and not the entire history. While this technique offers predictions, its efficacy is limited compared to our joint model, as shown in Figure 3.8. It illustrates the difference between true and predicted log-PSA levels, estimated with both exponential and joint models. The plot distinguishes predictions made with the exponential model (1647 predictions shown in blue) and our joint model (1671 predictions represented in black), underlying the limited predictive capacity of the exponential basic model, which handles 24 fewer data predictions due to the intrinsic structure of the prediction mechanism (points referring to time t_1 and t_2 can not be handled). Nevertheless, it is worth noting that the variability of the predicted values with the exponential model is higher than that with the proposed joint model. This highlights the superior predictive performance of the proposed joint model. Finally, the accuracy for PET-PSMA examination predictions, with the simple logistic model (threshold $\pi^* = 0.55$), as presented in Section 3.6.1, is 69.17%, which is improved by the 77.08% accuracy obtained with our joint model.

3.7 Final remarks and conclusions

Correct and quick identification of the locations of possible metastasis in prostate cancer is a challenging open problem. Despite the availability of several new techniques, their calibration is still debated. In particular, the results obtained using the sensitive nuclear examination known as PET-PSMA can be improved if correctly combined with a good estimation of the optimal time to perform the examination. The previous sections contain our proposal on how to estimate the optimal time to perform PET-PSMA which exploits information from the whole history data of each patient. We have introduced the joint model approach, addressing both PSA growth and the probability of a positive PET-PSMA, enriched with random effects to enable predictions for future patients. Our proposal is therefore not just a method for predicting individual PSA growth curve and time to PET-PSMA, but a proposal to drive optimal decisions regarding the patient, which is the core of personalized medicine. After explaining the model structure and the joint approach, we have discussed the optimal time estimation. Finally, the model has been estimated both on simulated and real data. Simulations were used to test the proposed model under challenging settings. In particular, we showed that the resurgence changing point time τ is difficult to estimate, as well as the regressive parameters entering the mean of the random effects. On the other hand, simulations also highlight the adaptability of the method to quite different growth patterns of PSA. The results obtained on real data, for an optimal probability $\pi^* = 0.55$ with a confidence of $\rho = 95\%$ on the result, give new insights into the model applicability and performance. Both the growth model estimated patterns and the logistic results are easy to interpret and in accordance with clinical evidence. Despite the improvements gained with our joint-model proposal, its complex mathematical structure makes it challenging the clinical daily applicability. Further research directions could include the implementation of a graphical interface to help clinicians exploit the model in an easy way: any new patient under analysis should be easily included in the database through the interface, enforcing the model accuracy, and enabling good and quick predictions.

Part II

Recurrent events data

Chapter 4

Survival analysis

Survival analysis is the branch of statistics that analyzes waiting times until an event of interest occurs. In medicine, usually, the events of interest are adverse events: infections, deaths, withdrawals from clinical trials, and many others. Survival analysis seeks to investigate the proportion of the population that will survive (or remain event-free) over time, the rate of events occurring, and the factors that can influence this event rate. More generally, survival analysis focuses on modeling time-to-event data.

4.1 Modeling Survival data

A survival dataset, for a population of n individuals, is generally composed of baseline covariates, time-varying covariates, and the time of the event (when it occurs) or the last observed time, when it does not occur (or something prevents from observing it) together with a dichotomous indicator variable (1 if event occurs, 0 otherwise). If the follow-up window is too short to observe the event of interest, we refer to the data as right-censored. Survival data are usually described and modeled in terms of survival and hazard probabilities.

The survival probability, usually indicated with $S(t)$, is the probability that an individual survives from the time origin t_0 to a specified future time t. The hazard is usually denoted by $\lambda(t)$ (or $h(t)$) and is the probability that an individual under observation at a time t will experiment an event in the infinitesimal time interval $[t, t + \Delta t)$: it represents the instantaneous event rate for an individual who survived until time t. Integrating the hazard function over $[0, t]$ gives the cumulative hazard function $\Lambda(t) = \int_0^t \lambda(s)ds$, which is linked to the survival by

$$S(t) = \exp\left(-\int_0^t \lambda(s)ds\right) = \exp(-\Lambda(t)). \tag{4.1}$$

Many models and methods are used to describe, evaluate and estimate survival, hazard, and cumulative hazard functions. In this chapter we present an overview of the most popular ones.

4.1.1 Estimating the Survival function

The survival probability can be non-parametrically estimated from observed survival data, using the Kaplan-Meier (KM) [KM58] and the Breslow methods (B) [Cox+50]. For a population under observation, composed of n individuals, suppose that H out of n patients experimented the event of interest during the follow-up period $[0, \tau]$ at ordered times $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$. In this particular scenario,

events are assumed to occur independently of one another, thus multiplying the probabilities of surviving from one interval to the next gives the cumulative survival probability. More formally, the probability of being alive at time $t_{(h)}$, namely $\hat{S}(t_h)$, is estimated from the probability $\hat{S}(t_{h-1})$ of being alive at $t_{(h-1)}$, the number n_h of patients alive at $t_{(h)}$, and the number d_h of events happening at $t_{(h)}$.
KM estimator uses the recursive formula

$$\hat{S}_{\text{KM}}(t_{(h)}) = \hat{S}_{\text{KM}}(t_{(h-1)})\left(1 - \frac{d_h}{n_h}\right), \quad \text{where } \hat{S}_{\text{KM}}(0) = 1, \tag{4.2}$$

which gives

$$\hat{S}_{\text{KM}}(t) = \prod_{h:t_{(h)}<t}\left(1 - \frac{d_h}{n_h}\right). \tag{4.3}$$

The step-function KM, a plot of the KM survival probability in time, provides a visual summary of the data.
Breslow method is based on a similar idea and suggests estimating the survival probability by $\hat{S}_{\text{B}}(t) = \exp(-\hat{\Lambda}(t))$, which gives

$$\hat{S}_{\text{B}}(t) = \prod_{h:t_{(h)}<t} e^{-d_h/n_h}, \quad \text{where } \hat{S}_{\text{B}}(0) = 1. \tag{4.4}$$

For small increments, the presented estimators are similar, since $e^{-x} \approx 1 - x$, which happens, in particular, when many subjects are still at risk. Thus, it can be shown for $n \to \infty$ that the estimators in Equations (4.3) and (4.4) are asymptotically equivalent.

The value of $\hat{S}_{\text{B}}(t)$ and $\hat{S}_{\text{KM}}(t)$ are constant between consecutive times of events, and therefore the estimated probability is a step function that changes value only at times of events.

Comparing Survival curves: the Log-Rank test

It is often of high clinical interest to assess whether there is any difference in survival probability (or cumulative incidence of events) among different groups of individuals. For example, in a clinical trial with a survival outcome, it might be interesting to compare survival probabilities between participants receiving a new drug and those receiving placebo (or standard therapy). In an observational study, the focus might be to compare survival probability between men and women, or between participants with and without a particular risk factor (e.g., hypertension or diabetes). In literature, several tests are available to compare survival among independent groups. The log-rank test is the most popular: it tests the null hypothesis of no difference in survival between two or more independent groups, compared to the alternative hypothesis of significant differences. The log-rank test compares the entire survival experience between groups and can be thought of as a test to determine whether the survival curves are identical (overlapping) or not. Survival curves are estimated for each group separately using the KM method, and then statistically compared. Considering two groups of individuals with $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$ times of events. Let's N_{1h} and N_{2h} be respectively the number of individuals at risk in $[t_{(h)}, t_{(h+1)})$, and O_{1h} and O_{2h} respectively the number of events in $[t_{(h)}, t_{(h+1)})$ for group 1 and group 2. Finally, $N_h = N_{1h} + N_{2h}$ and $O_h = O_{1h} + O_{2h}$.

The test comparing the survival probabilities $S_1(t)$ and $S_2(t)$ can be summarized by the following null and alternative hypotheses

$$\begin{aligned} H_0 &: S_1(t) = S_2(t), \\ H_1 &: S_1(t) \neq S_2(t). \end{aligned} \tag{4.5}$$

In both groups, under H_0, we have $O_{i,h} \sim Hypergeometric(N_h, N_{ih}, O_h)$, thus it comes by definition that

$$\begin{aligned} E_{ih} &= \mathbb{E}[O_{ih}] = O_h \frac{N_{ih}}{N_h}, \\ V_{ih} &= Var[O_{ih}] = E_{ih}\left(\frac{N_h - O_h}{N_h}\right)\left(\frac{N_h - N_{ih}}{N_h - 1}\right). \end{aligned} \tag{4.6}$$

For each time $t_{(h)}$, $h = 1, \ldots, H$, the log-rank test compares $O_{i,h}$ to its expectation under H_0, using the statistic

$$Z_i = \frac{\sum_{h=1}^{H}(O_{ih} - E_{ih})}{\sqrt{\sum_{h=1}^{H} V_{ih}}}, \tag{4.7}$$

which, under H_0, by central limit theorem (when H goes to infinity), converges in distribution to a standard normal distribution.

4.1.2 Estimating Hazard functions

Despite the survival function $S(t)$ and the hazard function $\lambda(t)$ being linked by Equation (4.1), it is not always easy to estimate $\lambda(t)$. A popular approach to estimating the hazard is to assume that the survival time follows a specific mathematical distribution. The most used distributions are reported in Table 4.1.

TABLE 4.1: Characteristics of the exponential, the Weibull, and the Gompertz distributions.

	Distributions		
Characteristic	Exponential	Weibull	Gompertz
Parameter	scale $\lambda > 0$	scale $\lambda > 0$ shape $\nu > 0$	scale $\lambda > 0$ shape $\alpha \in \mathbf{R}$
Hazard function	λ	$\lambda\nu t^{\nu-1}$	$e^{\alpha t}$
Cumulative hazard function	λt	λt^{ν}	$\frac{\lambda}{\alpha}(e^{\alpha t} - 1)$
Density function	$\lambda e^{-\lambda t}$	$\lambda\nu t^{\nu-1} e^{-\lambda t^{\nu}}$	$\lambda e^{(\alpha t)} e^{\frac{\lambda}{\alpha}(1-\exp(\alpha t))}$
Survival function	$e^{-\lambda t}$	$e^{-\lambda t^{\nu}}$	$e^{\frac{\lambda}{\alpha}(1-e^{\alpha t})}$

4.2 Cox model for time to event

One of the main focuses of survival analysis is to describe the relationship between survival times and covariates. Usually, this is done by a linear-like regression model for the log hazard. The Cox model is the most famous semi-parametric model with this regression structure. It links the covariates $Z_i = (z_{1i}, ..., z_{pi})^T$ of patient i to the individual log hazard as

$$\log \lambda_i(t) = \alpha_0(t) + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi}, \tag{4.8}$$

or, in exponential form,

$$\lambda_i(t) = \lambda_0(t) e^{\alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi}}, \text{ where } \lambda_0(t) = e^{\alpha_0(t)}. \tag{4.9}$$

The label semi-parametric refers to the hazard structure, obtained as a product of a non-parametric function $\lambda_0(t)$ and a parametric term $e^{\alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi}}$, where $\lambda_0(t) = e^{\alpha_0(t)}$. This model is also called proportional-hazard model, as the ratio of individual hazards does not depend on time: taking individual i and individual j and their hazard $\lambda_i(t)$ and $\lambda_j(t)$, the ratio is then

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) e^{\alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi}}}{\lambda_0(t) e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}} = \frac{e^{\alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi}}}{e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}}. \tag{4.10}$$

This means that, even if the baseline hazard $\lambda_0(t)$ remains unspecified, the Cox model can be estimated by the partial-likelihood method. The partial likelihood function for Cox model was introduced by Cox et al. [Cox72] and can be written as

$$PL(\alpha_1, \ldots, \alpha_p) = \prod_{i=1}^{n} \left\{ \frac{\lambda_i(t)}{\sum_{j:t_j<t_i} \lambda_j(t)} \right\}^{\delta_i} = \prod_{i=1}^{n} \left\{ \frac{e^{\alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi}}}{\sum_{j:t_j<t_i} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}} \right\}^{\delta_i}, \tag{4.11}$$

where $1 - \delta_i$ is an indicator function, equal to one when the patient event Is not observed due to right censoring, and t_j is the time of event for patient j. From the log-partial-likelihood

$$\begin{aligned} \log(PL(\alpha_1, \ldots, \alpha_p)) &= \sum_{i=1}^{n} \delta_i \left\{ \log \lambda_i(t) - \log \sum_{j:t_j<t_i} \lambda_j(t) \right\} \\ &= \sum_{i=1}^{n} \delta_i \left\{ \alpha_1 z_{1i} + \alpha_2 z_{2i} + \ldots + \alpha_p z_{pi} - \log \sum_{j:t_j<t_i} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}} \right\}, \end{aligned} \tag{4.12}$$

it is possible to derive the system of score functions U_{α_k}, for all $k = 1, \ldots, p$,

$$U_{\alpha_k} = \frac{\partial \log(PL(\alpha_1, \ldots, \alpha_p))}{\partial \alpha_k} = \sum_{i=1}^{n} \delta_i \left\{ z_{ki} - \frac{\sum_{j:t_j<t_i} z_{kj} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}}{\sum_{j:t_j<t_i} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}} \right\}, \tag{4.13}$$

and the estimator $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)^T$ by solving the system $(U_{\alpha_1}, \ldots, U_{\alpha_p}) = 0$. It can be shown that the solution $\hat{\alpha}$ is consistent and asymptotically normally distributed with mean α.

The Cox model can also be extended to include time-varying covariates, as

$$\lambda_i(t) = \lambda_0(t) e^{\alpha_1 z_{1i}(t) + \alpha_2 z_{2i}(t) + \cdots + \alpha_p z_{pi}(t)}, \tag{4.14}$$

and all the theory explained still holds.

4.2.1 Residuals

The Cox model main assumptions can be summarized as

- the baseline hazard rate $\lambda_0(t)$ is common between all patients at any time t, meanings that all individuals are assumed to experience the same baseline hazard;
- the effect of the regression variables on the instantaneous hazard experienced by an individual is assumed to remain constant over time;
- the regression coefficients α does not vary with time.

It is possible to see Equations (4.11), (4.12), and (4.13) from a different perspective, that allows constructing a fully parametric model for survival analysis. Let us consider again n individuals, for each individual i the p-dimensional vector of covariate $Z_i = (z_{1i}, \ldots, z_{pi})^T$ is collected. Let D be the set of indices of the non-censored individuals and let R_i be the set of indices of those who were under observation when the i-th individual experienced the event, namely $R_i = \{j : t_j < t_i\}$. Then, for each $i \in D$, the probability P_m of sample index $m \in R_i$ is

$$P_m = \frac{e^{\alpha_1 z_{1m} + \alpha_2 z_{2m} + \cdots + \alpha_p z_{pm}}}{\sum_{j \in R_i} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}}, \tag{4.15}$$

In this scenario, Z_i is considered to be a random vector with components conditional expectations

$$\mathbb{E}[z_{ki}|R_i] = \frac{\sum_{j \in R_i} z_{kj} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}}{\sum_{j \in R_i} e^{\alpha_1 z_{1j} + \alpha_2 z_{2j} + \cdots + \alpha_p z_{pj}}}, \quad k = 1, ..., p. \tag{4.16}$$

Substituting Equation (4.16) in Equation (4.13), we obtain this equivalent formulation for the score functions that we use to estimate $\alpha = (\alpha_1, ..., \alpha_p)^T$:

$$U_{\alpha_k} = \sum_{i \in D} \{z_{ki} - \mathbb{E}[z_{ki}|R_i]\}. \tag{4.17}$$

Denoting with $\hat{\alpha}$ the estimate of the parameter of interest, and with $\hat{\mathbf{E}}[z_{ki}|R_i]$ the estimate of the true expectation $\mathbf{E}[z_{ki}|R_i]$ computed by sobstitution of $\hat{\alpha}$, we can define the residual $\hat{r}_i = (\hat{r}_{1i}, ..., \hat{r}_{pi})^T$ as

$$\hat{r}_{ki} = z_{ki} - \hat{\mathbf{E}}[z_{ki}|R_i], \quad k = 1, ..., p. \tag{4.18}$$

It can be proved that the $\hat{r}_i$ residuals are uncorrelated [Sch82]. If the proportional hazard assumption holds, then $\mathbf{E}[\hat{r}_i] \simeq 0$, and a plot of $\hat{r}_{ik}$ versus t_i should be centered around zero. A formal test, named Schoenfeld Residual Tests, can also be conducted, with proportional assumption holding as a null hypothesis. For further details, we refer to [Sch82; PH15].

4.2.2 Simulate survival times from Cox proportional hazards models

Simulating the event times for a Cox model is quite useful in many different contexts. To better understand the mechanism of the simulation procedure we first describe the constant-covariate case, then the time-varying covariate framework, which is of main interest. In particular, combining Equations (4.1) and (4.9), we can generally define for a patient i the distribution function of the Cox model as

$$\begin{aligned} F_i(t) &= 1 - \exp(-\int_0^t \lambda_0(s)\, e^{\sum_{j=1}^{p} \alpha_j z_{ji}} ds) \\ &= 1 - \exp(-\Lambda_0(t) e^{Z_i^T \alpha}), \end{aligned} \tag{4.19}$$

and explot it to generate the time of event t as $t = \Lambda_0^{-1}[-\log(u)\exp(-Z_i^T\alpha)]$, where u is a sample from a uniform variable $u \sim U(0,1)$, as demonstrated by both Leemis [Lee87] and Bender et al. [BAB05].

Things become more complicated when also time-varying covariates are included in the model. Many different cases can be analyzed and treated differently depending on the structure of the time-varying covariates. We focus here only on discrete time-varying covariates with multiple but finite changes in the follow-up period, while for ohter senarios we refer to [Aus13].

We consider here, for the sake of simplicity, a model with $p = 2$, where for each patient i a baseline covariate z_{1i} and a time-varying covariate $z_{2i}(t)$ are measured, thus the array of the individual covariates for patient i is $z_i(t) = (z_{1i}, z_{2i}(t))^T$. The hazard Equation (4.14) becomes

$$\lambda_i(t) = \lambda_0(t) e^{\alpha_1 z_{1i} + \alpha_2 z_{2i}(t)}. \tag{4.20}$$

We are interested in the case when the time-varying covariate has a finite number G of changes, respectively at times $0 = g_{0,i} < g_{1,i} < g_{2,i} < \cdots < g_{G,i}$, which may differ from patient to patient. Defining the time-varying covariate as

$$z_{2i}(t) = \begin{cases} v_{1i}, & 0 \le t < g_{1,i}, \\ v_{2i}, & g_{1,i} \le t < g_{2,i}, \\ \quad \vdots & \\ v_{Gi}, & g_{G-1,i} \le t < g_{G,i}, \end{cases} \tag{4.21}$$

we can rewrite the hazard as a step-wise constant function

$$\lambda_i(t) = \begin{cases} \lambda_0(t) e^{\alpha_1 z_{1i} + \alpha_2 v_{1i}}, & 0 \le t < g_{1,i}, \\ \lambda_0(t) e^{\alpha_1 z_{1i} + \alpha_2 v_{2i}}, & g_{1,i} \le t < g_{2,i}, \\ \quad \vdots & \\ \lambda_0(t) e^{\alpha_1 z_{1i} + \alpha_2 v_{Gi}}, & g_{G-1,i} \le t < g_{G,i}, \end{cases} \tag{4.22}$$

and decompose the problem of generating the time-to-event for a Cox hazard model with a time-varying covariate, to the problem of generating the time-to-event for a Cox hazard model with a fixed covariate, on G intervals. The ease of the procedure only depends on the shape of $\lambda_0(t)$. Here we consider the case $\lambda_0(t) = \lambda_0$, just as an example, and derive the full methodology to simulate the time-to-event.

The cumulative hazard can be derived by integrating Equation (4.22), to obtain

$$H_i(t) = \begin{cases} \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{1i}}(t), & 0 \leq t < g_{1,i}, \\ \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{1i}}(g_{1,i}) + \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{2i}}(t - g_{1,i}), & g_{1,i} \leq t < g_{2,i}, \\ \quad \vdots & \\ \sum_{j=1}^{G-1} \left[\lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{ji}}(g_{j,i} - g_{j-1,i})\right] + \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{Gi}}(t - g_{G-1,i}), & g_{G-1,i} \leq t < g_{G,i}, \end{cases} \tag{4.23}$$

and then inverting each of the piece-wise components of the cumulative hazard function, we have the inverse cumulative hazard function

$$H_i^{-1}(t) = \begin{cases} \dfrac{t}{\lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{1i}}}, & t \in R_1, \\ \dfrac{t - H(g_{1,i})}{\lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{2i}}} + g_{1,i}, & t \in R_2, \\ \quad \vdots & \\ \dfrac{t - H(g_{G-1,i})}{\lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{Gi}}} + g_{G-1,i}, & t \in R_G, \end{cases} \tag{4.24}$$

where the intervals can be defined as

$$\begin{aligned} R_1 &= [0, \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{1i}} g_{1,i}], \\ R_2 &= [H(g_{1,i}), H(g_{1,i}) + \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{2i}}(g_{2i} - g_{1,i})], \\ &\vdots \\ R_G &= [H(g_{G-1,i}), H(g_{G-1,i}) + \lambda_0 e^{\alpha_1 z_{1i} + \alpha_2 v_{Gi}}(g_{G,i} - g_{G-1,i})]. \end{aligned} \tag{4.25}$$

To simulate a survival time for individual i it suffices to evaluate $H_i^{-1}(-\log u)$, where the value of $\log u$ determines which of the G components of the inverse is used.

4.3 Geneneralized linear models for time to event data

Generalized linear models (GLM) can be used as a parametric tool to model survival data. The approach is different from the hazard-based methods proposed so far. To better understand how to apply GLM, we should think of the data to be organized in the time-person format: the dataset is composed of a row for each patient at each time $\{t_k\}_{k=1}^K$ which is in the time-partition-grid. Times t_k do not need to correspond to the event time $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$. Furthermore, let D_k be an auxiliary indicator variable, such that $D_k = 1$ if $T \leq t_k$ and 0 otherwise. The survival probability at time k is $\mathbb{P}(T > k)$, or equivalenty $\mathbb{P}(D_k = 0)$ and can be factorized as

$$\begin{aligned} \mathbb{P}(T > k) &= \prod_{j=1}^{k} \mathbb{P}(T > t_j | T > t_{j-1}) \\ &= \prod_{j=1}^{k} \mathbb{P}(D_j = 0 | D_{j-1} = 0) \end{aligned} \tag{4.26}$$

In particular, if $\Delta_k = t_k - t_{k-1}$ is small enough, $\mathbb{P}(T > t_{k+1}|T > t_k)$ is the discrete-time hazard function. An easy way to parametrically approximate the hazard $\mathbb{P}(T > t_{k+1}|T > t_k)$ is to construct a logistic regression model. Hernan and Robins [HR20] proposed to model the logit of the hazard as

$$\begin{aligned}\text{logit}[\mathbb{P}(D_{k+1} = 1|D_k = 0, z_1, \ldots, z_n)] = \theta_{0,k} + \theta_1 z_i + \theta_2 z_i \times k + \theta_3 z_i \times k^2,\\ \theta_{0,k} = \theta_0 + \theta_4 \times k + \theta_5 \times k^2.\end{aligned} \tag{4.27}$$

It should be noted that the time-varying intercept and the product between covariates and time model a time-varying hazard and ratio, respectively. When using the hazard ratio as a measure of causal effect, an important property of the hazard ratio needs to be taken into account: because the hazards vary over time, the hazard ratio generally does too. That is, the ratio at time t_k may differ from that at time t_{k+1}. The validity of this procedure requires no misspecification of the hazards model. Of course, the smaller the time intervals $\Delta_j = t_k - t_{k-1}$ are, the better the statistical methods will approximate the survival. To show how the logistic approximation holds, let us consider a simple scenario with a simple one-dimensional binary covariate $z \in \{0, 1\}$ (e.g., the treatment indicator). The discrete-time hazard ratio is then

$$\frac{\mathbb{P}(D_{k+1} = 1|D_k = 0, z = 1)}{\mathbb{P}(D_{k+1} = 1|D_k = 0, z = 0)} = \exp(\alpha_1), \tag{4.28}$$

meaning that for all k we have

$$\mathbb{P}(D_{k+1} = 1|D_k = 0, z_i = 1) = \mathbb{P}(D_{k+1} = 1|D_k = 0, z_i = 0) \times \exp(\alpha_1). \tag{4.29}$$

Taking the logarithm on both sides we obtain

$$\begin{aligned}\log(\mathbb{P}(D_{k+1} = 1|D_k = 0, z_i = 1)) &= \log(\mathbb{P}(D_{k+1} = 1|D_k = 0, z_i = 0)) \times (\alpha_1)\\ &= \alpha_{0,k} + \alpha_1.\end{aligned} \tag{4.30}$$

If the hazard at $k + 1$ is close to 0 (i.e., $\mathbb{P}(D_{k+1} = 1|D_k = 0, z) \simeq 0$), then the hazard is approximately equal to the odds

$$\mathbb{P}(D_{k+1} = 1|D_k = 0, z) \simeq \frac{\mathbb{P}(D_{k+1} = 1|D_k = 0, z)}{\mathbb{P}(D_{k+1} = 0|D_k = 0, z)}, \tag{4.31}$$

thus

$$\log \frac{\mathbb{P}(D_{k+1} = 1|D_k = 0, z)}{\mathbb{P}(D_{k+1} = 0|D_k = 0, z)} = \text{logit}[\mathbb{P}(D_{k+1} = 1|D_k = 0, z)] \simeq \alpha_{0,k} + \alpha_1. \tag{4.32}$$

That is, if the hazard is close to zero at k + 1, we can approximate the log hazard ratio α_1 by θ_1 in a logistic model as $\text{logit}[\mathbb{P}(D_{k+1} = 1|D_k = 0, z)] = \theta_0, k + \theta_1$ like the one we used in Equation (4.27). As a rule of thumb, the approximation is often considered to be accurate enough when $\mathbb{P}(D_{k+1} = 1|D_k = 0, z) < 0.1$ for all k. This rare event condition can almost always be guaranteed to hold taking a time unit k that is short enough.

4.4 Example and Code

We destinate this section to clarify the literature introduced in Chapter 4 by means of a toy example, applying the methods and models presented, and reporting the related R code. We simulate a dataset to emulate a randomized trial that aims at estimating the treatment efficacy of a new therapy for cancer compared with the standard of care. The main focus is to understand if patients who received the new treatment have longer expected life compared to those receiving the standard of care. The dataset, simulated in R, is composed of 200 patients, for each of them the baseline covariates measured are

- `trt`: 0 = standard of care, 1 = new treatment
- `time`: time of death
- `status`: censoring status
- `age`: age of the patient at enrollment
- `sex`: male or female
- `biomarker`: an indicator of tumor progression

The interest is to analyze the time-to-censoring to understand which covariates are associated with it, and to understand if the treatment is improving the life-duration. The dataset structure is in Table 4.2. Patients are followed from enrollment (time 0) until the end of follow-up (time 200). The status variable indicates whether the patient was alive or not at the time of recording. Status 1 refers to death, status 0 indicates that the patient did not die before the end of follow-up. The dataset is in the person-time format, meaning that it is composed of a single line for each patient.

TABLE 4.2: Some rows of the dataset under analysis.

	trt	time	status	age	sex	biomarker
1	1	200	0	65	0	1.77
2	0	200	0	58	0	2.18
3	0	174	1	72	0	2.55
4	1	200	0	67	1	1.80
5	1	123	1	69	1	1.91
6	0	197	1	61	1	1.87

TABLE 4.3: Some rows of the dataset analyzed in the long format.

	trt	age	sex	biomarker	tstart	time	status
1	1	65	0	1.77	0	1	0
2	1	65	0	1.77	1	2	0
3	1	65	0	1.77	2	3	0
4	1	65	0	1.77	3	4	0
5	1	65	0	1.77	4	5	0
6	1	65	0	1.77	5	6	0

Cox estimating function (`coxph`) provided by `R` package `survival` [The+] can be applied to this format but the dataset needs to be modified to apply the GLM model (with `glm` function in `R` package `stats` [R C]). For this reason, the `long.dataset` function splits the dataset over all the times selected (line 5), in this case the smallest unit is considered, and creates a record for each person-time.

```
1  ## Function to modify the dataset in the long format
2  long.dataset<-function(data)
3  {
4    # Define times to split dataset
5    events=1:max(data$time)
6
7    # Split data over event-times
8    data.long <- survSplit(Surv(time,status) ~ .,data=data, cut=events)
9
10   # add variable time^2 to the dataset
11   data.long$startsq=data.long$tstart^2
12
13   # return dataset in time-person format
14   return(data.long)
15 }
```

The new long-format dataset is in Table 4.3, where `time` indicates the time at which each record is collected, and `tstart` indicates the last time at which each patient was observed (at last occurence recorded in the previous row). All the other variables remain the same.

The R package `survival` [The+] provides the `coxph` function to fit the Cox model for time-to-event. The function requires the user to specify the event/status variable, the corresponding time and the covariates to be included in the analysis. We present here two different Cox models, namely `cox0` and `cox1` (lines 1-6), which differ for the treatment covariate and compare them with the ANOVA test. In addition, we investigate the treatment effect on the survival function using the log-rank-test (line 10). We finally test the proportional-hazard assumption using Schoenfeld's residuals, looking graphically at them and performing the corresponding test which shows that Cox proportional model assumption is satisfied (lines 12–13). The `biomarker` covariate is significantly related to survival times if treatment is not considered, but is no more significant when treatment is added. To understand the reason behind this, further models should be tested, namely adding the interaction between treatment and biomarker value, but this is out of the scope in this section.

```
1  # Cox model without terapy
2    cox0=coxph(Surv(time,status) ~ age + sex + biomarker,
3   data = dataset.cancer.long)
4  # Cox model with terapy
5    cox1=coxph(Surv(time,status) ~ trt + age + sex + biomarker,
6   data = dataset.cancer.long)
7  # comparison with ANOVA
8    anova(cox0,cox1,test="Chisq")
9  # Log-rank-test
10   survdiff(Surv(time,status) ~ trt, data = dataset.cancer)
11 # Analysis of residuals
12   residuals(cox1)
13   print(cox.zph(cox1))
```

We list in Table 4.4 to Table 4.8 the outputs, used to interpret the results.

TABLE 4.4: Output of Model `cox0`: coefficients estimates with standard error and pvalues.

	term	estimate	std.error	statistic	p.value
1	age	-0.01	0.01	-1.18	0.24
2	sex	-0.15	0.17	-0.89	0.37
3	biomarker	0.21	0.09	2.33	0.02

TABLE 4.5: Output of Model `cox1`: coefficients estimates with standard error and pvalues.

	term	estimate	std.error	statistic	p.value
1	trt	-0.49	0.18	-2.73	0.01
2	age	-0.01	0.01	-1.30	0.19
3	sex	-0.23	0.17	-1.34	0.18
4	biomarker	0.14	0.10	1.44	0.15

TABLE 4.6: ANOVA test output to compare model 1 (`cox1`) and model 2 (`cox0`).

	loglik	chisq	df	p.value
1	-1196.57			
2	-1200.36	7.58	1	0.0059

TABLE 4.7: Log-rank test output (p.value = 0.004).

	trt	N	obs O	exp E	$(O-E)^2/E$	$(O-E)^2/V$
1	0	105.00	80.00	63.24	4.44	8.15
2	1	95.00	61.00	77.76	3.61	8.15

TABLE 4.8: Analysis of Residuals.

	chisq	df	p.value
trt	3.49	1	0.06
age	0.76	1	0.38
sex	1.25	1	0.26
biomarker	0.30	1	0.59
GLOBAL	4.81	4	0.31

TABLE 4.9: Output of model `glm0` (AIC: 1692.2), with coefficient estimates, standard errors and pvalues.

	estimate	std. error	statistic	p.value
(Intercept)	-8.0778	0.7535	-10.72	0.0000
tstart	0.0433	0.0092	4.69	0.0000
I(tstart^2)	-0.0001	0.0000	-3.12	0.0018
age	-0.0070	0.0084	-0.83	0.4075
sex	-0.0841	0.1711	-0.49	0.6230
biomarker	0.1433	0.0867	1.65	0.0981

As a second approach, we model the simulated data using the binomial generalized linear model on the dataset in the long-time-person format, as

```
## GLM model
    glm0=glm(status ~ tstart+I(start^2) + age + sex + biomarker ,
                family=binomial(),
                data=dataset.cancer.long)
    glm1=glm(status ~ tstart+I(start^2) + age + sex + biomarker + trt ,
                family=binomial(),
                data=dataset.cancer.long)
    anova(glm0,glm1,test="Chisq")
```

The GLM models, with the ANOVA test, confirm that survival time is strongly dependent on the treatment received. Tables 4.9 and 4.10 report the output.

Finally, we give a small graphical comparison of the Cox and GLM model to describe time-to-event data (see Figure 4.1). We reduce the models of interest to the treatment variable only, to make it more easily representable. We first construct the GLM model and use it to make predictions, time by time, for treated and standard-of-care arms. We then use bootstrap to obtain 95% confidence survival curve. We plot the KM estimates with 95% confidence interval for the Cox model and the median GLM curves with 95% confidence region, exploiting `survfit` and `ggsurvplot` `R` functions, respectively in `survival` and `survminer` package [The+; Kas+].

```
## Model comparison : Cox an GLM
  glmcompare=glm(status ~ tstart+I(start^2) + trt ,
                family=binomial(),
                data=dataset.cancer.long)
# Make predictions for treated and untreated
  times=seq(1,max(dataset.cancer.long$time)+20,1)
  LT=length(times)

  dataset0=as.data.frame(cbind(rep(0,LT),c(0,times[1:(LT-1)]),times))
  colnames(dataset0)=c("trt","tstart","time")
  glmpred=1-predict(glmcompare,dataset0,type='response')
  datasetplot_trt0=as.data.frame(cbind(times,cumprod(glmpred)))
        colnames(datasetplot_trt0)=c("timepred", "pred")

  dataset1=as.data.frame(cbind(rep(1,LT),c(0,times[1:(LT-1)]),times))
  colnames(dataset1)=c("trt","tstart","time")
  glmpred=1-predict(glmcompare,dataset1,type='response')
  datasetplot_trt1=as.data.frame(cbind(times,cumprod(glmpred)))

  colnames(datasetplot_trt1)=c("timepred", "pred")
```

TABLE 4.10: Output of model `glm1` (AIC: 1687.3), with coefficient estimates, standard errors and pvalues.

	estimate	std. error	statistic	p.value
(Intercept)	-7.7223	0.7652	-10.09	0.0000
tstart	0.0440	0.0092	4.76	0.0000
I(tstart^2)	-0.0001	0.0000	-3.18	0.0015
age	-0.0078	0.0084	-0.93	0.3542
sex	-0.1384	0.1721	-0.80	0.4214
biomarker	0.0856	0.0918	0.93	0.3511
trt	-0.4634	0.1779	-2.60	0.0092

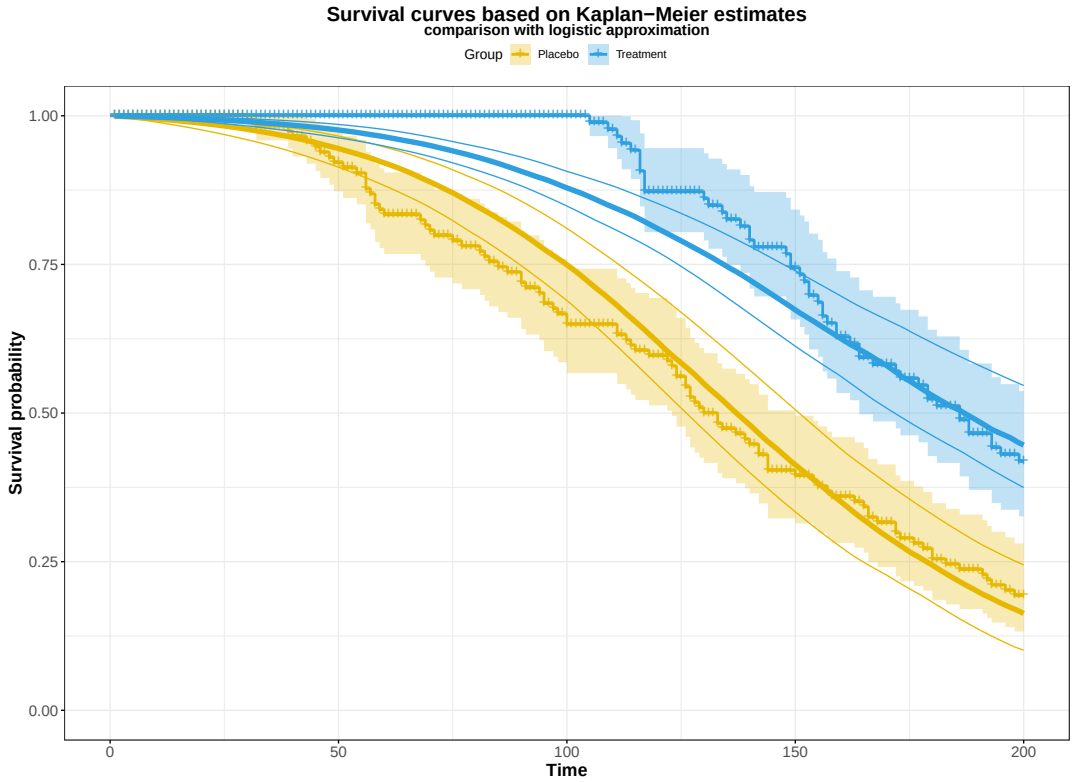

FIGURE 4.1: Comparison survival curves estimated with KM (step solid-point line with 95% filled interval) and logistic bootstrapped median (bold solid lines with 95% confidence lines).

4.5 Censoring problem

Statistical censoring is a crucial concept in statistical modeling, particularly when dealing with time-to-event data in survival analysis. Various types of censoring exist, depending on the statistical quantity of interest. These include right censoring, left censoring, and interval censoring, each posing unique challenges and requiring specific modeling approaches. Understanding and appropriately addressing these assumptions are essential for ensuring the reliability of statistical analyses and the validity of conclusions drawn from survival data.

In the survival framework, right censoring occurs when the exact time of the event of interest is not observed or recorded within the study period, which may happen for various reasons, such as the study concluding before an event happens or participants being lost to follow-up. Three main assumptions are usually considered to describe right censoring mechanism. The first assumption, named **independent censoring**, is the most useful of the three to draw correct inferences that compare the survival experience of two or more groups. Independent censoring means that, within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk, with respect to their survival experience. In other words, censoring is independent given that it is random within any subgroup of interest (defined by observed and measured covariates). The second assumption, called **random censoring**, is a stronger and more restrictive assumption than independent censoring. Random censoring means that subjects censored at time t are representative of all the study subjects who remained at risk at that time with respect to their survival experience. In other

words, the failure rate for censored subjects is assumed to be equal to the failure rate for uncensored subjects who remained in the risk set. Note that independent censoring is random censoring, conditional on each level of covariates. Finally, **non-informative** censoring occurs if the distribution of survival times provides no information about the distribution of censoring times, and vice versa. The assumption of non-informative censoring is usually easily justifiable when censoring is independent and/or random; nevertheless, these three assumptions are not equivalent. In many situations, censoring can be reasonably considered non-informative, and hence standard likelihood-based procedures can be used without the necessity of any correction. However, in other situations, it is not clear whether censoring is non-informative; in fact, it is sometimes clear that censoring is related to survival times. Despite its crucial importance, the non-informative censoring assumption is not possible to test without making additional restrictions, for example, restrictions on the joint distribution of times of events and censoring times. Choosing the appropriate assumption is crucial, as violations can lead to biased estimates and erroneous conclusions. Researchers must carefully assess the assumptions underlying these mechanisms to ensure the reliability and validity of their statistical models. For a detailed description with additional examples, we refer to [KK12].

4.5.1 Right dependent censoring for survival analysis: how to apply IPCW

So far we assumed all data to be fully recorded or to be randomly censored (independent censoring mechanism implying no dependence between lifetime T_i and censoring time C_i for each patient i).
Being $h_i(t)$ the censoring mechanism, the assumption we made so far is

$$h_i(t|T_i, z_{1i}, ..., z_{pi}) = h_i(t). \tag{4.33}$$

However, more than often, subjects lost to follow-up are not representative of the patients at risk. They may have characteristics that differ from the average patient characteristics: in this case, we refer to the censoring as dependent censoring, which implies

$$h_i(t|T_i, z_{1i}, ..., z_{pi}) \neq h_i(t). \tag{4.34}$$

When Equation (4.34) holds, meaning that time-to-event and censoring are dependent, the methods presented so far will poorly perform and give biased estimates. In particular, when clinicians believe that withdrawal is likely to be related to health status, as the covariate that best determines the health status is also one of the covariates used in the definition of the event time, corrections should be applied while estimating survival probability and time-to-event distribution. In causal inference literature, the most popular approach to deal with missing data problems in survival analysis is to apply the Inverse Probability Censoring Weigthing (IPCW) technique [Wil+18]. IPCW estimator corrects for censored subjects by giving extra weight to subjects who are not censored. Each subject i is weighted by the inverse of an estimate of the conditional probability of having remained uncensored until time t. To illustrate how IPCW works an example will be given. Let's consider a sample of 4 individuals and suppose that the estimated chance of remaining uncensored until time t = 5 is 1/4, hence 3 out of each 4 subjects are censored before time t = 5. For each subject at risk, there would have been 3 extra subjects at risk for the event of

interest at time t = 5 if censoring was absent. By weighting the contribution of subjects that are not censored at time t = 5 by P^{-1}(survive to time $t = 5$)=4, the censored subjects are included in the final estimates.

We present here the KM-IPCW estimator, indicated by $\hat{S}_{KM-IPCW}(t)$, but IPCW methods can be combined with many other approaches. Combining Equation (4.3) with IPCW, we obtain

$$\hat{S}_{KM-IPCW}(t) = \prod_{j:t_j<t} \left(1 - \frac{\sum_{i:\delta_i(t_j)=1} \hat{W}_i(t_j)}{\sum_{i=1}^{n} Y_i(t_j)\hat{W}_i(t_j)}\right), \tag{4.35}$$

with $Y_i(t_j)$ at-risk indicator function for patient i at time t_j, and

$$\hat{W}_i(t) = \frac{1}{\hat{S}_C(t|z_{1i}, ..., z_{pi})} \tag{4.36}$$

the inverse of the estimated probability of censoring $\hat{S}_C(t|z_{1i}, ..., z_{pi})$ happening after time t, estimated for example using KM estimator.

To give a small insight on how to apply and evaluate KM-IPCW we show a toy simulated example, based on [Wil+18]. We consider a dataset composed of $n = 100$ patients. For each patient i the collected data are:

- the patient indicator `id`;
- the level of the continuous biomarker of interest $\texttt{Z}_1$ measured at baseline;
- the treatment received $\texttt{Z}_2$ (1 or 0);
- the `time`=min(t_i, c_i) which is the minimum between the event time t_i and the withdrawal time c_i;
- the withdrawal indicator `delta`=$I(c_i < t_i)$;
- the event indicator `status`.

Thus, the resulting dataset contains entries like in Table 4.11.

To understand if the assumption in Equation (4.33) holds, we perform a survival analysis for time-to-withdrawal using Cox model:

```
CoxModel <- coxph(Surv(time, delta) ~ Z1+Z2, data = dataset)
```

and we obtain that withdrawal is significantly dependent on covariates $\texttt{Z}_1$ and $\texttt{Z}_2$ (p-values are almost zero, respectively $6.51 \cdot 10^{-6}$ and $1.43 \cdot 10^{-7}$). Table 4.12 summarizes the results.

TABLE 4.11: Some rows of the dataset under analysis.

	id	Z1	Z2	tstart	time	delta	status
1	1	-0.63	0	0	317	0	1
2	2	0.18	0	0	550	1	0
3	3	-0.84	1	0	171	1	0
4	4	1.60	0	0	211	1	0
5	5	0.33	0	0	187	1	0
6	6	-0.82	1	0	1	1	0

From the analyses performed so far, it is necessary to apply IPCW correction to obtain an unbiased estimate of survival probability (see following code). To compute IPCW, the dataset must be in the time-person format (rows 1–9). Lines 12–13 defines the censoring mechanism used to evaluate the weights (line 15–27).

```
1  # Modifing the dataset from person format to time-person format
2  test.long <- survSplit(dataset, cut=dataset$time, end="time",
3                         start="start", event="status")
4  test.long <- test.long[order(test.long$id,test.long$time),]
5  test.long.cens <- survSplit(dataset, cut=dataset$time, end="time",
6                              start="start", event="delta")
7  test.long.cens <- test.long.cens[order(test.long.cens$id,
8                                         test.long.cens$time),]
9  test.long$event <- test.long.cens$delta
10
11 # Construct model for censoring
12 C0 <- coxph(Surv(start, time, censored) ~ 1, data = test.long)
13 CZ <- coxph(Surv(start, time, censored) ~ Z1+Z2, data = test.long)
14
15 # Compute the weigths
16 C0fit <- summary(survfit(C0), times = test.long$start)
17 test.long$K0ti <- C0fit$surv
18 test.long$KZti <- NULL
19 for(i in 1:nrow(test.long))
20 {
21   datai <- test.long[i,]
22   sfiCZ <- survfit(CZ, newdata = datai)
23   ssfiCZ <- summary(sfiCZ, times = datai$start)
24   test.long$KZti[i] <- ssfiCZ$surv
25 }
26 test.long$WUnStab <- 1/test.long$KZti
27 test.long$WStab <- test.long$K0ti/test.long$KZti
```

Finally, we exploit the IPCW technique to obtain a correction to KM estimator in lines 1–4 (in the following code), and compare the result plotting the survival curves in Figure 4.2. We evaluate both weigths and stabilized weigths (see [HR20], Technical point 12.2). Once the weigths are evaluated, we can compute the survival probability, applying the IPCW correction:

```
1 # Estimating time-to-event with KM, KM-IPCW, and KM-IPCW Stabilized
2 KM<-survfit(Surv(start, time , status) ~ Z2, data = test.long)
3 KM_IPCW <-survfit(Surv(start, time , status) ~ Z2,data = test.long,
    weights = WUnStab)
4 KM_IPCWStab <-survfit(Surv(start, time , status) ~ Z2, data = test.long
    , weights = WStab)
```

TABLE 4.12: Output of Model `CoxModel`, with coefficient estimates, standar errors and pvalues.

	term	estimate	std.error	statistic	p.value
1	Z1	0.85	0.19	4.51	0.00
2	Z2	1.98	0.38	5.26	0.00

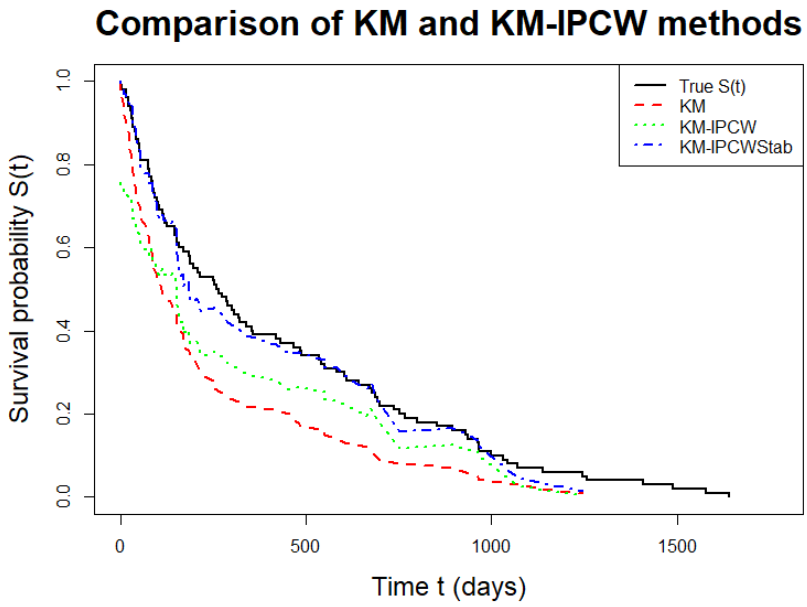

FIGURE 4.2: Survival probability curves evaluated with KM (red), KM-IPCW (green), and KM-IPWStab (blue) for a time-to-event setting with informative censoring. Results are compared with true survival curve (solid black line).

For further details on methods and implementation of IPCW, we refer to [Wil+18].

Chapter 5

Background of recurrent events analysis

In science and technology, interest often lies in studying processes that generate events repeatedly over time. Such processes are referred to as recurrent event processes and the data they provide are called recurrent event data. Recurrent event data arise in fields such as medicine and public health, business and industry, reliability, social sciences, and insurance. In these settings, the main interests include (i) understanding and describing (individual) event processes, (ii) identifying and characterizing variation across a population of processes, (iii) comparing groups of processes, and (iv) determining the relationship with fixed external agents (e.g., baseline covariates), and time-varying factors (e.g., time-dependent covariates) to event occurrence [CL07].

In medicine, whenever the same patient can be involved in more than one event, of the same type or not, we refer to it as a recurrent event setting. Examples of recurrent events can be epileptic seizures, heart attacks, tumor resurgence, infections, and many more. A common characteristic among these events is the intrinsic correlation between those occurring in the same subject. The statistical literature on the analysis of recurrent events for medicine has grown rapidly over the past twenty years and a variety of models and methods has been developed. This chapter provides a little sum-up of the background of recurrent events. We first introduce the notation to describe relevant models, explain their underlying assumptions and properties, consider settings where they are appropriate, and briefly discuss how to fit these models. Parametric, nonparametric, and semiparametric methods are briefly discussed. Modeling recurrent events can be approached in several ways: the most exploited ones are through event counts or waiting times. We particularly focus on one-type-of-events methods based on counts and rate functions, starting from the Poisson parametric model, relaxing the parametric structure, and describing the semi-parametric Andersen-Gill model [AG82], which is the main focus of the second part of the thesis. We finally discuss the applicability of the Poisson model on a simplified example, which is later highly discussed in Chapter 6.

5.1 Notation and Framework

In a recurrent event process, with starting point set to $t = 0$, the ordered event times are $T_1 < T_2 < \cdots < T_{n_e}$, where T_j is the time of event j, and n_e is the total number of events. It can be defined the associated counting process $\{N(t), 0 \leq t\}$, which stores

the cumulative number of events that occurred in $[0, t]$ as

$$N(t) = \sum_{j=1}^{\infty} I(T_j \leq t). \tag{5.1}$$

Counting processes are usually defined to be right continuous, namely $N(t) = N(t^+)$, where t^+ (and t^-) denotes a time that is infinitesimally greater (smaller) than t. The instantaneous probability of an event occurring at time t, conditional on the process history $X(t^-)$ in $[0, t)$, is given by the intensity function

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(N(t + \Delta t) - N(t) | X(t^-))}{\Delta t}, \tag{5.2}$$

or equivalently

$$\lambda(t)dt = \mathbb{E}(dN(t) | X(t^-)). \tag{5.3}$$

The right-continuity property makes Equations (5.2) and (5.3) well defined.

Sometimes, it is reasonable to assume that the intensity depends on some measured baseline or time-dependent covariates. We will present, in this chapter, some possible dependencies of intensity functions on the internal process history, while always assuming covariates to be external. In particular, we define $Z(t)$ to be the history over $[0, t]$ of external covariates (baseline and time-dependent), and we define the history of the whole process $X(t)$ to be $X(t) = \{N(s),\ Z(s) : 0 < s < t\}$. Finally, for each recurrent event process observed, the at-risk indicator function $Y(t)$ can be defined as

$$Y(t) = I(0 < t < \tau), \quad \text{with } [0, \tau] \text{ the observation period considered,} \tag{5.4}$$

and Equation (5.2) is sometimes rewritten as

$$\lambda(t)Y(t)dt = \mathbb{E}(dN(t) | X(t^-)). \tag{5.5}$$

This function defines, time by time, if the process can account for a new event or not: when dealing with multiple process settings we assume only one process can jump at time.

For a recurrent event process, with the intensity in Equation (5.2) and Markovian structure (see [CL07]), the probability of experimenting $n_e > 0$ events at times $t_1 < t_2 < \cdots < t_{n_e}$ over the time interval of interest $[0, \tau]$ is

$$\prod_{j=1}^{n_e} \lambda(t_j | X(t_j)) \exp\left(- \int_0^{\tau} \lambda(u | X(u)) du \right). \tag{5.6}$$

Moreover, the probability of no events occurring in $[s, t]$ is

$$\mathbb{P}\big(N(t) - N(s) = 0 | N(s+)\big) = \exp\left(- \int_s^t \lambda(u | X(u)) du \right). \tag{5.7}$$

Using Chapman-Kolmogorov equations [She93], Equation (5.7), can be extended to obtain the probability

$$\mathbb{P}\big(N(t) = n_e | N(0) = 0\big) = \int_0^t \lambda(\omega)\mathbb{P}\big(N(\omega) = n_e - 1 | N(0) = 0\big) \times \exp\bigg(-\int_\omega^t \lambda(v|H(v))dv\bigg)d\omega. \tag{5.8}$$

and the density functions of the interarrival times $\Delta_x = T_x - T_{x-1}$ to be

$$f_{\Delta_x|T_{x-1}}(u_x|t_{x-1}) = \lambda(t_{x-1} + u_x)\exp\bigg(\int_{t_{x-1}}^{t_{x-1}+u_x} \lambda(\omega)d\omega\bigg). \tag{5.9}$$

The theory presented in Equations (5.6)–(5.9) refers to a Markovian setting with a time-dependent intensity function $\lambda(t)$, but, in some particular circumstances, it may be of interest to assume the intensity function $\lambda_x(t)$ also depends on the number of events x already accounted as

$$\mathbb{P}\big(N(t + \Delta t) - N(t) = 1 | N(t) = x\big) = \lambda_x(t)\Delta t + o(\Delta t), \quad \text{when } \Delta t \to 0. \tag{5.10}$$

The notation introduced so far holds for a single recurrent event process. When we deal with multiple recurrent event processes (e.g., one for each patient under analysis), we will add the subscript i to refer to a specific one.
For further details on recurrent events theory, we refer to [CL07].

5.2 Parametric model: Poisson model

The Poisson process is the most famous model to describe counting data. It has several equivalent definitions, we report here the one based on the intensity function. Poisson intensity function is of the form

$$\lambda(t|X(t)) = \rho(t), \tag{5.11}$$

meaning that the probability of an event occurring in a time interval may depend on it but not on the process and covariate history $X(t)$. In particular, this structure implies that the following two properties hold:

1. $N(t) - N(s) \sim Poisson(\int_s^t \rho(u)du)$;
2. if $(s_1, t_1]$ and $(s_2, t_2]$ are non overlapping intervals, then $N(s_1) - N(t_1)$ and $N(s_2) - N(t_2)$ are independent random variable.

If the function $\rho(t)$ is time-dependent, the process is called inhomogeneous, if it is constant then it is called homogeneous. The intensity function is often modified to include effect of the process of covariates history $Z(t)$, to obtain the multiplicative-Poisson-intensity-model

$$\lambda(t|X(t)) = \rho_0(t;\alpha)\exp(\beta Z(t)). \tag{5.12}$$

This formulation is particularly useful to deal with a population of n patients, each described by their own Poisson process, with a slightly different shape, that can be explained by a common baseline parametric function $\rho_0(t;\alpha)$ with parameter α, but different covariates. We report here how to apply the maximum likelihood method

to estimate the parameters of the Poisson processes for n individuals, and, to this purpose, we introduce the subscript i to refer to the i-th individual of the population. Each individual process is described by its intensity, with structure given by Equation (5.12), where $\rho_0(t;\alpha)$ depends on a $r \times 1$ parameter α, and β is a $p \times 1$-vector of coefficients (p being the dimension of covariates process $Z_i(t)$). Finally, $\theta = (\alpha, \beta)$ the vector of coefficients to estimate. Individuals are assumed to be independent of each other, given the covariates, thus the likelihood function is the product of the individual likelihood terms $L_i(\theta)$:

$$L_i(\theta) = \prod_{j=1}^{n_i} \rho_0(t_{ij};\alpha)\exp(\beta Z_i(t_{ij}))\exp\left(-\int_0^\tau Y_i(s)\rho_0(s;\alpha)\exp(\beta Z_i(s))ds\right), \quad (5.13)$$

where n_i is the number of events of individual i at times $(0 < t_{i1}, \ldots, t_{in_i} < \tau)$. Full likelihood $L(\theta)$ and log-likelihood $l(\theta)$ are

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta), \quad (5.14)$$

$$l(\theta) = \sum_{i=1}^{n} \int_0^\tau Y_i(s)[\log \rho_i(s;\alpha)dn_i(s) - \rho_i(s;\theta)ds], \quad (5.15)$$

with $\rho_i(s;\alpha) = \rho_0(s;\alpha)\exp(\beta Z_i(s))$ for brevity. Deriving the log-likelihood $l(\theta)$ with respect to the parameters of interest, namely α and β, leads to the partial scores functions

$$\begin{aligned} U_\alpha &= \sum_{i=1}^{n} \int_0^\tau Y_i(s)\frac{\partial \log \rho_0(s;\theta)}{\partial \alpha}[dn_i(s) - \rho_i(s;\theta)ds] = 0, \\ U_\beta &= \sum_{i=1}^{n} \int_0^\tau Y_i(s)Z_i(s)[dn_i(s) - \rho_i(s;\theta)ds] = 0, \end{aligned} \quad (5.16)$$

which compose the full score vector $U(\theta) = (U_\alpha^T(\theta), U_\beta^T(\theta))^T$. The explicit MLE estimator $\hat{\theta}$, which is the solution of $U(\theta) = 0$, is not always explicit, and sometimes it is necessary to exploit optimization tools to maximize $l(\theta)$ and find $\hat{\theta}$.

In the easiest case, namely when the n Poisson processes are homogeneous with constant-intensity-function $\rho_i(t;\theta) = \rho$ and no covariate dependency is modeled, the maximization problem of the log-likelihood function $l(\theta)$ has the explicit solution:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} \int_0^\tau Y_i(s)dn_i(s)}{n\tau} = \frac{\sum_{i=1}^{n} n_i(\tau)}{n\tau}. \quad (5.17)$$

Moreover, if we assume the population to be divided into m subgroups (e.g., based on the treatment received), each having its own constant-intensity-function ρ_j, for groups $j = 1, \ldots, m$, then

$$\hat{\rho}_j = \frac{\sum_{i\in\text{group } j} \int_0^\tau Y_i(s)dn_i(s)}{m_j\tau} = \frac{\sum_{i\in\text{group } j} n_i(\tau)}{m_j\tau}, \quad j = 1, \ldots, m. \quad (5.18)$$

where m_j is the number of individuals in group j, such that $n = \sum_{j=1}^{m} m_j$. A common assumption is to divide the individual under observation based on the treatment

received and to compare the respective intensity functions to estimate treatment efficacy.

It can be proved that estimators in Equations (5.17) and (5.18) are unbiased, as it holds

$$\mathbb{E}[\hat{\rho}] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} n_i(\tau)}{n\tau}\right] = \frac{\sum_{i=1}^{n} \mathbb{E}[n_i(\tau)]}{n\tau} = \rho, \tag{5.19}$$

$$\mathbb{E}[\hat{\rho}_j] = \mathbb{E}\left[\frac{\sum_{i\in\text{group } j} n_i(\tau)}{m_j\tau}\right] = \frac{\sum_{i\in\text{group } j} \mathbb{E}[n_i(\tau)]}{m_j\tau} = \rho_j, \quad j = 1, \ldots, m, \tag{5.20}$$

from Poisson model properties (1) and (2). Many other parametric models for recurrent events have been studied in the last 30 years. For a full detailed description, we refer to [CL07].

5.3 Non-Parametric and Semi-Parametric models

For the sake of realism, sometimes, the assumptions made on the baseline hazard function $\rho_0(t;\alpha)$ are relaxed: to obtain a more flexible scenario the baseline function is not assumed to have any particular parametric form, reducing to the more general formulation of $\rho_0(t)$. Infer recurrent events parameters for semi-parametric and non-parametric models presents more challenges than for the parametric ones. We present the main results reported in literature, starting from a simpler non-parametric case. Let's consider the case of n individuals, each having $n_i(t)$ events described by an independent Poisson process with intensity $\rho(t)$. The interest is inferring the expected value of $n_i(t, t+dt)$, which is equal to $d\mu(t) = \rho(t)dt$.
For this purpose, we start by evaluating the estimating function U

$$U = \sum_{i=1}^{n} Y_i(s)[dn_i(s) - d\mu(s)] = 0, \tag{5.21}$$

and deriving the estimator $d\hat{\mu}(s)$ as

$$d\hat{\mu}(s) = \frac{\sum_{i=1}^{n} Y_i(s)dn_i(s)}{\sum_{i=1}^{n} Y_i(s)}. \tag{5.22}$$

It can be easily show that $d\hat{\mu}(s)$ in an unbiased estimator

$$\begin{aligned}\mathbb{E}(d\hat{\mu}(s)) =& \mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i(s)dn_i(s)}{\sum_{i=1}^{n} Y_i(s)}\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i(s)dn_i(s)}{\sum_{i=1}^{n} Y_i(s)}\middle| Y_1(s), \ldots, Y_n(s)\right]\right] \\ =& d\mu(s).\end{aligned} \tag{5.23}$$

To finally obtain the estimate of interest for $\mu(t)$, we should compute $\mu(t) = \int_0^t d\mu(s)$ by

$$\hat{\mu}(t) = \int_0^t d\hat{\mu}(s) = \int_0^t \frac{\sum_{i=1}^{n} Y_i(s)dn_i(s)}{\sum_{i=1}^{n} Y_i(s)} = \sum_{h:t_{(h)}\le t} \frac{\sum_{i=1}^{n} Y_i(t_{(h)})dn_i(t_{(h)})}{\sum_{i=1}^{n} Y_i(t_{(h)})}, \tag{5.24}$$

where $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$ denote the H ordered distinct event times across the n individuals.

The more interesting model for recurrent-event data for our purpose, which is the main focus of this second part of the thesis, is the semi-parametric Poisson model, also called the Andersen-Gill model [AG82], and presented in the next section.

5.3.1 The Andersen and Gill model

The semi-parametric Poisson model identified by the individual intensity function formulation

$$\rho_i(t|Z(t)) = Y_i(t)\rho_0(t)\exp(\beta Z_i(t)), \tag{5.25}$$

where $Z(t)$ is the full population process history, $Z_i(t)$ the individual one. Equation (5.25) takes the name of the Andersen-Gill model [AG82]. For this semi-parametric scenario, the inference is much more complicated: we report here the profile likelihood and the partial likelihood approach that are equivalent and commonly used to estimate the parameter β involved.

Similarly to the non-parametric case presented in Equations (5.21)–(5.24), we treat $d\mu_0(d) = \rho_0(s)ds$ and we can replace the score functions in Equation (5.16) by

$$\begin{aligned} U_1 &= \sum_{i=1}^{n} Y_i(s)[dn_i(s) - \exp(\beta Z_i(s))d\mu_0(s)] = 0, \qquad 0 \le s, \\ U_2 &= \sum_{i=1}^{n} \int_0^{\tau} Y_i(s)Z_i(s)\left[dn_i(s) - \frac{\sum_{l=1}^{n} Y_l(s)dN_l(s)}{\sum_{l=1}^{n} Y_l(s)\exp(\beta Z_l(s))d\mu_0(s)}\exp(\beta Z_i(s))\right] = 0. \end{aligned} \tag{5.26}$$

Solving the system of Equation (5.26) leads to the estimators $\hat{\mu}_0(s)$ and $\hat{\beta}$.
Analogously, the formula for U_2 in Equation (5.26) can be derived working with the partial likelihood $PL(\beta)$ with expression

$$PL(\beta) = \prod_{h=1}^{H}\left[\frac{\exp(\sum_{i\in S_h}\beta Z_i(t_{(h)})\beta)}{[\sum_{l=1}^{n} Y_l(t_{(h)})\exp(\beta Z_l(t_{(h)}))]^{s_h}}\right], \tag{5.27}$$

where H are the ordered events time $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$ and S_h the set of s_h individuals having an event at time $t_{(h)}$, for $h = 1, \ldots, H$. Note that in theory, for continuous-time problems, s_h should be equal to one, but we assume here event times to be recorded with finite precision. Taking the partial likelihood in Equation (5.27) and differentiating it with respect to β leads to the same estimating formulation for U_2 as reported in Equation (5.26).

5.4 Case study: MLE for censored Poisson data

We briefly report here a motivational real case study, that is the main focus of the work we fully describe in Chapter 6. We exploit this example, simplifying it a bit, to clarify and apply some of the theoretical results presented in this chapter and to make some small mathematical considerations, before going into details in Chapter 6.

The example we present is based on a real clinical trial conducted by GSK to develop a vaccine for Chronic Obstructive Pulmonary Disease (COPD), characterized by respiratory adverse events. The vaccine aims at reducing the individual number of these critical events related to this pathology, but a correct estimation of the efficacy can be prevented by informative withdrawals: the higher the number of events the higher the probability of dropout. All the details about the study can be found in Chapter 6 and in the published papers [And+22; Aro+22].

The recurrent event histories of each patient of a randomized population, either to standard of care or to an experimental treatment, are recorded as i.i.d. copies of the initial segments of two-point processes. In the simplest case, the two processes can be modeled as i.i.d. copies of two homogeneous Poisson processes with intensities λ_S and λ_T (S standard of care, T treatment), so that the Intensity Ratio (IR)

$$\text{IR} = \frac{\lambda_T}{\lambda_S}, \tag{5.28}$$

or its logarithm, is an inverse measure of treatment effect; the greater IR is, the worse is the treatment since we are dealing with negative events. In the case of therapeutic vaccines, an opposite but equivalent direct measure of treatment effect is Vaccine Efficacy

$$\text{VE} = 1 - \frac{\lambda_T}{\lambda_S}. \tag{5.29}$$

In the presence of informative dropout, the reduction of the observation time may create biases in estimating IR and, consequently, VE; the goal of this example is to explore simple computable models to describe the data. Suppose each patient is scheduled to be observed over a deterministic follow-up time FU, but may be subject to right-censoring by a censoring random variable C, so that the actual observation Time is

$$T = \min\{FU, C\}. \tag{5.30}$$

In particular, for each patient i, let introduce the following random quantities: the number of events N_i, happening at times $T_{i,1}, \ldots, T_{i,N_i}$, and the final time $T_{i,N_i+1} = \min\{FU, C_i\}$, where C_i is the censoring time.

For each patient $i = 1, \ldots, n$, let us record the observed values for these quantities, in the following statistics:

- n_i the number of events occurred to the $i-$th patient in the observed time-window;
- $t_{1,i}, ..., t_{i,n_i}$ the observed times of events;
- the last observation times is $t_{i,n_i+1} = FU$ if the i-th patient is not censored, otherwise $t_{i,n_i+1} = c_i$, where FU is a deterministic follow-up time and c_i is the censoring time ;
- δ_i, the indicator of censoring, which equals 0 if $t_{i,n_i+1} = FU$, and equals 1 if $t_{i,n_i+1} = c_i$;
- τ_i, the treatment label (1 if the patient is in the treatment group, 0 if control).

To model the censoring mechanism, i.e., the dropout, in a parametric way, several choices are available, we present here, as an example, two simple possibilities: the

first one is modeling the censoring independently from the events, the second one is considering hazard of withdrawal to be proportional to the intensity of the events.

5.4.1 Poisson model with independent censoring

The easiest model to apply is a Poisson-process model for recurrent events with the independent censoring assumption for withdrawal. In the case of independent censoring, $C_1, \ldots, C_n$ are assumed to be i.i.d random variable as

$$C_1, \ldots, C_n \overset{\text{i.i.d.}}{\sim} Exponential(\mu). \tag{5.31}$$

Then the likelihood is as follows:

$$\begin{aligned} L(\lambda_t, \lambda_s, \mu) = & \prod_{i:\delta_i=0} \lim_{\Delta\to 0} \frac{\mathrm{P}\left(\bigcap_{k=1}^{n_i} T_{i,k} \in (t_{i,k} \pm \Delta), \delta_i = 0\right)}{\Delta^{n_i}} \times \\ & \prod_{i:\delta_i=1} \lim_{\Delta\to 0} \frac{\mathrm{P}\left(\bigcap_{k=1}^{n_i} T_{i,k} \in (t_{i,k} \pm \Delta), T_{i,n_i+1} \in (t_{i,n_i+1} \pm \Delta), \delta_i = 1\right)}{\Delta^{n_i+1}} \\ = & \prod_{\substack{i:\delta_i=0,\\ \tau_i=0}} \exp(-\lambda_S\, t_{i,n_i+1}) \lambda_S^{n_i} \exp(-\mu\, t_{i,n_i+1}) \prod_{\substack{i:\delta_i=1,\\ \tau_i=0}} \exp(-\lambda_S\, t_{i,n_i+1}) \lambda_S^{n_i} \mu \exp(-\mu\, t_{i,n_i+1}) \times \\ & \prod_{\substack{i:\delta_i=0,\\ \tau_i=1}} \exp(-\lambda_T\, t_{i,n_i+1}) \lambda_T^{n_i} \exp(-\mu\, t_{i,n_i+1}) \prod_{\substack{i:\delta_i=1,\\ \tau_i=1}} \exp(-\lambda_T\, t_{i,n_i+1}) \lambda_T^{n_i} \mu \exp(-\mu\, t_{i,n_i+1}), \end{aligned} \tag{5.32}$$

so that the loglikelihood is

$$l(\lambda_T, \lambda_s, \mu) = l(\mu) - \lambda_S \sum_{i:\tau_i=0} t_{i,n_i+1} + \log(\lambda_S) \sum_{i:\tau_i=0} n_i - \lambda_T \sum_{i:\tau_i=1} t_{i,n_i+1} + \log(\lambda_T) \sum_{i:\tau_i=1} n_i, \tag{5.33}$$

and we can ignore the terms in μ to evaluate the MLE esimator fro λ_S and λ_T. The Maximum Likelihood Estimator (MLE) of IR is, as expected,

$$\widehat{\mathrm{IR}}_{\mathrm{MLE}} = \frac{\widehat{\lambda_T}}{\lambda_S} = \frac{\sum_{i:\tau_i=1} n_i \sum_{i:\tau_i=0} t_{i,n_i+1}}{\sum_{i:\tau_i=1} t_{i,n_i+1} \sum_{i:\tau_i=0} n_i}, \tag{5.34}$$

i.e., the ratio of the two average numbers of events per unit of time.

The model implementation in `R` is shown in the following code. In the code, 2000 different datasets, composed of 500 patients each, are simulated. In case of no censoring, the mean numbers of events, respectively in the control and treatment arms are 10 and 5. The mean time to withdraw is 70 days and the follow-up period is 100 days. The true ratio is 0.5, while the mean ratio is estimated to be 0.51 with a standard deviation of 0.15.

```
set.seed(1)
N=500                                  # population size
FU=rep(100,N)                          # follow-up period
MU=0.70                                # hazard of censoring
    LAMBDA_S=0.10                          # intensity for standard of
   care
LAMBDA_T=0.05                          # intensity for treatment
VE=1-(LAMBDA_T/LAMBDA_S)               # true vaccine efficacy
LAMBDARAPPORTO=(LAMBDA_T/LAMBDA_S)     # ratio of rates
```

```
MC=2000                                  # number of Monte Carlo repetitions
ESTIMATES=array(MC)
EVENTS=array(N)


for(mc in 1:MC)
{
 TRT=sample(c(0,1),N,replace = T) # randomization arm
 CENS=rexp(N,MU)                  # censoring time
 TMAX=ifelse(CENS<FU,CENS,FU)     # time of observation
 IDcens=ifelse(CENS<FU,1,0)       # indicator of censoring
 for(p in 1:N)
 {
  EVENTSTIME=cumsum(rexp(100,(TRT[p]*LAMBDA_T+(1-TRT[p])*LAMBDA_S)))
  EVENTSTIME_SELECTED=EVENTSTIME[which(EVENTSTIME<TMAX[p])]
  EVENTS[p]=length(EVENTSTIME_SELECTED)
 }

 ESTIMATES[mc]=(sum(EVENTS[TRT==1])*sum(TMAX[TRT==0]))/(sum(EVENTS[TRT
    ==0])*sum(TMAX[TRT==1]))
}
```

5.4.2 Poisson model with proportional censoring hazard

We can then complicate the model, by assuming dependent censoring. The simplest form of dependences we can think about is when the censoring hazard censoring variable is proportional to events intensity:

$$C_i \sim \begin{cases} Exponential(\kappa\lambda_S), & \text{if } \tau_i = 0, \\ Exponential(\kappa\lambda_T), & \text{if } \tau_i = 1. \end{cases} \tag{5.35}$$

The parameter κ quantifies the informativeness of the dropout mechanism. The likelihood can be written

$$\begin{aligned} L(\lambda_t, \lambda_s, k) = &\prod_{i:\delta_i=0} \lim_{\Delta\to 0} \frac{\mathrm{P}\left(\bigcap_{k=1}^{n_i} T_{i,k} \in (t_{i,k} \pm \Delta), \delta_i = 0\right)}{\Delta^{n_i}} \times \\ &\prod_{i:\delta_i=1} \lim_{\Delta\to 0} \frac{\mathrm{P}\left(\bigcap_{k=1}^{n_i} T_{i,k} \in (t_{i,k} \pm \Delta), T_{i,n_i+1} \in (t_{i,n_i+1} \pm \Delta), \delta_i = 1\right)}{\Delta^{n_i+1}} \\ = &\prod_{\substack{i:\delta_i=0,\\ \tau_i=0}} \exp(-\lambda_S t_{i,n_i+1}) \lambda_S^{n_i} \exp(-\kappa\lambda_S t_{i,n_i+1}) \prod_{\substack{i:\delta_i=1,\\ \tau_i=0}} \exp(-\lambda_S t_{i,n_i+1}) \lambda_S^{n_i} \kappa\lambda_S \exp(-\kappa\lambda_S t_{i,n_i+1}) \times \\ &\prod_{\substack{i:\delta_i=0,\\ \tau_i=1}} \exp(-\lambda_T t_{i,n_i+1}) \lambda_T^{n_i} \exp(-\kappa\lambda_T t_{i,n_i+1}) \prod_{\substack{i:\delta_i=1,\\ \tau_i=1}} \exp(-\lambda_T t_{i,n_i+1}) \lambda_T^{n_i} \kappa\lambda_T \exp(-\kappa\lambda_T t_{i,n_i+1}), \end{aligned} \tag{5.36}$$

so that the loglikelihood is

$$\begin{aligned} l(\lambda_t, \lambda_s, k) = &- \lambda_S(1+\kappa) \sum_{i:\tau_i=0} t_{i,n_i+1} + \log(\lambda_S)\Big(\sum_{i:\tau_i=0} n_i + \sum_{\substack{i:\delta_i=1,\\ \tau_i=0}} 1\Big) + \kappa \sum_{\substack{i:\delta_i=1,\\ \tau_i=0}} 1 + \\ &- \lambda_T(1+\kappa) \sum_{i:\tau_i=1} t_{i,n_i+1} + \log(\lambda_T)\Big(\sum_{i:\tau_i=1} n_i + \sum_{\substack{i:\delta_i=1,\\ \tau_i=1}} 1\Big) + \kappa \sum_{\substack{i:\delta_i=1,\\ \tau_i=1}} 1, \end{aligned} \tag{5.37}$$

where $\sum_{i:\delta_i=1,\tau_i=0} 1$ is the number of censored observations under S and $\sum_{i:\delta_i=1,\tau_i=1} 1$ is the number of censored observations under T. The MLE of IR is easily proven to be

$$\widehat{\mathrm{IR}}_{\mathrm{MLE}} = \frac{\widehat{\lambda_T}}{\lambda_S} = \frac{(\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1) \sum_{i:\tau_i=0} t_{i,n_i+1}}{\sum_{i:\tau_i=1} t_{i,n_i+1} (\sum_{i:\tau_i=0} n_i + \sum_{i:\delta_i=1,\tau_i=0} 1)}, \tag{5.38}$$

where

$$\widehat{\lambda_T} = \frac{(\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1)}{\sum_{i:\tau_i=1} t_{i,n_i+1}}, \quad \widehat{\lambda_S} = \frac{(\sum_{i:\tau_i=0} n_i + \sum_{i:\delta_i=1,\tau_i=0} 1)}{\sum_{i:\tau_i=0} t_{i,n_i+1}}, \tag{5.39}$$

which shows the interesting effect that censored observations are counted as extra events in both arms.

Statement 5.4.1. *The MLE estimator $\widehat{IR}_{MLE}$ is consistent as*

$$\widehat{IR}_{MLE} \quad \overset{n\to\infty}{\to} \quad IR\,. \tag{5.40}$$

Proof.

$$\begin{aligned} \mathbb{E}[\widehat{\lambda}_T] &= \mathbb{E}\left[\frac{\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}} \right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}} \middle| \{t_{i,n_i+1}\}_{\tau_i=1} \right] \right] \\ &= \mathbb{E}\left[\frac{\sum_{i:\tau_i=1} \mathbb{E}[n_i | t_{i,n_i+1}]}{\sum_{i:\tau_i=1} t_{i,n_i+1}} + \frac{\sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}} \right] \\ &= \lambda_T + \mathbb{E}\left[\frac{\sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}} \right] \\ &= \lambda_t + k\lambda_T, \end{aligned} \tag{5.41}$$

where it can be shown that the process which counts the number of withdrawals, respectively in treatment and placebo arm, is a Poisson process with intensity $k\lambda_T$ and $k\lambda_S$. Similarly, we can derive

$$\mathbb{E}[\widehat{\lambda}_S] = \lambda_S + k\lambda_S. \tag{5.42}$$

Moreover it can be computed the variance as

$$\begin{aligned} Var[\widehat{\lambda}_T] &= Var\left[\frac{\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}}\right] \\ &= Var\left[\mathbb{E}\left[\frac{\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}}\middle| \{t_{i,n_i+1}\}_{\tau_i=1}\right]\right] + \\ &\quad \mathbb{E}\left[Var\left[\frac{\sum_{i:\tau_i=1} n_i + \sum_{i:\delta_i=1,\tau_i=1} 1}{\sum_{i:\tau_i=1} t_{i,n_i+1}}\middle| \{t_{i,n_i+1}\}_{\tau_i=1}\right]\right] \\ &= \mathbb{E}\left[\frac{\sum_{i:\tau_i=1} Var[n_i|t_{i,n_i+1}]}{(\sum_{i:\tau_i=1} t_{i,n_i+1})^2}\right] \\ &= \mathbb{E}\left[\frac{\lambda_T \sum_{i:\tau_i=1} t_{i,n_i+1}}{(\sum_{i:\tau_i=1} t_{i,n_i+1})^2}\right] \leq \frac{\lambda_T}{\sum_{i:\tau_i=1} FU} \overset{n\to\infty}{\to} 0. \end{aligned} \tag{5.43}$$

and, with the same procedure,

$$Var[\widehat{\lambda}_S] \overset{n\to\infty}{\to} 0. \tag{5.44}$$

This means that $\widehat{\lambda}_T \overset{\mathbb{P}}{\to} \lambda_T$ and $\widehat{\lambda}_S \overset{\mathbb{P}}{\to} \lambda_S$, and

$$\mathbb{E}\left[\frac{\widehat{\lambda_T}}{\lambda_S}\right] \overset{n\to\infty}{\underset{\mathbb{P}}{\to}} \frac{\mathbb{E}[\widehat{\lambda}_T]}{\mathbb{E}[\widehat{\lambda}_S]} = \frac{\lambda_T}{\lambda_S}, \tag{5.45}$$

meaning that the estimator $\widehat{\text{IR}}_{\text{MLE}}$ is consistent. □

The memoryless structure of the Poisson process also guarantees that the naive estimator

$$\widehat{\text{IR}}_{\text{NAIVE}} = \frac{\widehat{\lambda_T}}{\lambda_S} = \frac{\sum_{i:\tau_i=1} n_i \sum_{i:\tau_i=0} t_{i,n_i+1}}{\sum_{i:\tau_i=1} t_{i,n_i+1} \sum_{i:\tau_i=0} n_i} \tag{5.46}$$

is a consistent estimator for IR (the proof is similar to the one we reported for $\widehat{\text{IR}}_{\text{MLE}}$).

From simulations, we can clearly show that for finite and small sample size $\widehat{\text{IR}}_{\text{MLE}}$ outperforms $\widehat{\text{IR}}_{\text{NAIVE}}$, at it exploits more information contained in the dataset. The model implementation, in R, is the following.

```
set.seed(1)
n=500                                  # population size
FU=rep(10,n)                           # follow-up period
LAMBDA_S=0.4                           # intensity of standard of care
LAMBDA_T=0.1                           # intensity of treatment
K=1                                    # rate constant of censoring
VE=1-(LAMBDA_T/LAMBDA_S)               # true vaccine efficacy
LAMBDARAPPORTO=(LAMBDA_T/LAMBDA_S)     # ratio of rates

MC=2000                                # number of Monte Carlo repetitions
ESTIMATES=array(MC)
ESTIMATES_NAIVE=array(MC)
EVENTS=array(N)

for(mc in 1:MC)
{
 TRT=sample(c(0,1),n,replace = T)                  # randomization arm
 CENS=rexp(n,K*(TRT*LAMBDA_T+(1-TRT)*LAMBDA_S))    # censoring time
 TMAX=ifelse(CENS<FU,CENS,FU)                      # time of observation
 IDcens=ifelse(CENS<FU,1,0)                        # indicator of
    censoring
```

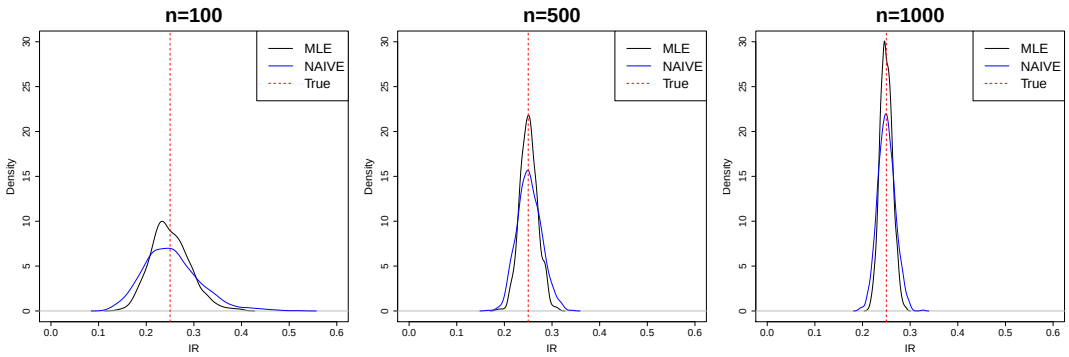

FIGURE 5.1: $\widehat{\text{IR}}_{\text{MLE}}$ (black) and $\widehat{\text{IR}}_{\text{NAIVE}}$ (blue) distributions for 1000 repetitions with increasing sample size $n = 100, 500, 1000$. Red vertical line is on the true value.

```
21
22  for(p in 1:N)
23 {
24   EVENTSTIMEALL=cumsum(rexp(500,(TRT[p]*LAMBDA_T+(1-TRT[p])*LAMBDA_S)))
25   EVENTSTIME=EVENTSTIMEALL[EVENTSTIMEALL<TMAX[p]]
26   EVENTS[p]=length(EVENTSTIME)
27  }
28   N0=sum(EVENTS[TRT==0])+sum(IDcens[TRT==0])
29   N1=sum(EVENTS[TRT==1])+sum(IDcens[TRT==1])
30   ESTIMATES[mc]=(N1*sum(TMAX[TRT==0]))/(N0*sum(TMAX[TRT==1]))
31   ESTIMATES_NAIVE[mc]=(sum(EVENTS[TRT==1])*sum(TMAX[TRT==0]))/(sum(
       EVENTS[TRT==0])*sum(TMAX[TRT==1]))
32 }
```

The code reported above simulates 2000 different datasets, composed of 500 patients each. In case of no censoring, the mean numbers of events, respectively in the control and treatment arms are 4 and 1. The mean time to withdraw is, respectively for placebo and treatment, 2.5 and 10 days and the follow-up period is 10 days. The withdrawal percentage is around 80%. The ratio IR is 0.25. We repeated the estimating procedure for $n = 100, 500, 1000$ shown in Figure 5.1.

Chapter 6

Estimating in practice vaccine efficacy in randomized recurrent event trials with informative censoring

This Chapter is based on the paper:
Amongero, Martina and Callegaro, Andrea and Gasparini, Mauro and Moraschini, Luca and Costantini, Marco and Vansteelandt, Stijn. **Estimating in practice vaccine efficacy in randomized recurrent event trials with informative censoring: an application to COPD setting.** (Work in progress).

Although vaccine efficacy estimation for single endpoints has been extensively studied in various settings and applications, extending this methodology to recurrent events settings is not straightforward, and the complexity of the problem can increase substantially. This study focuses on estimating vaccine efficacy for recurrent infection settings with informative censoring, particularly when the outcome itself contributes to the censoring event's informativeness. The goal is to provide a clear and practical guideline on applying the Andersen and Gill (AG) model, one of the most widely used models for recurrent events, in conjunction with inverse probability of censoring weighting method (IPCW), a causal inference technique used to correct biased estimates. The method is applied to a real case study, an investigational clinical trial conducted by GSK to develop a vaccine for Chronic Obstructive Pulmonary Disease. The vaccine aims at reducing the individual number of moderate and severe events (Acute Exacerbations) related to this pathology, but a correct estimation of the efficacy can be prevented by informative withdrawals: the higher the number of events the higher the probability of dropout.

We briefly describe the available literature in Section 6.1, and we introduce the motivating case study which justifies our work in Section 6.2. We then give a detailed explanation of the method studied (Section 6.3) and test it on some simulated datasets (Section 6.4). We finally apply the method to the real case study in Section 6.5 and give an overall conclusion in Section 6.6.

Conflict of interest
Luca Moraschini, Marco Costantini, and Andrea Callegaro are employees of the GSK group of companies. Marco Costantini and Andrea Callegaro own shares in the GSK group of

companies. Martina Amongero is a Ph.D. student at Politecnico di Torino with a scholarship sponsored by the GSK group of companies.

Funding
This work was sponsored and financially supported by GlaxoSmithKline Biologicals SA.

Data Availability statement
Anonymised individual participant data and study documents can be requested for further research from `www.clinicalstudydatarequest.com`.

6.1 Literature

Recurrent events problems are common in many areas. In medicine, typical examples are epileptic seizures, heart attacks, tumor resurgence, infections, ..., which all can occur more than once in a single patient. Statistical analyses which aim to relate to occurrence of such events to subject-specific covariates, are typically complicated by the intrinsic correlation that is present between events occurring in the same subject. A large literature was developed to accommodate this (see [Aka+18] for an overview), with solutions that can largely be classified into 3 classes. Some methods model correlation between recurrent events under a Markov assumption that future events depend only on the immediate past. Well-known examples are semi-parametric methods based on the Andersen and Gill model [AG82] and the Prentice-Williams-Peterson (PWP) model [PWP81], which all invoke a proportional hazards assumption. Other approaches model the dependency between recurrent events via shared random effects in so-called joint frailty models [TD04]. A third class of approaches refrains from making explicit assumptions about the correlation structure, by viewing it as a nuisance parameter in a semi-parametric marginal model; marginal models instead take it into account, during the estimation procedure, using a sandwich variance estimator. Poisson and Negative binomial cases are the most famous models which refer to the latter category [Jah08; CKS19].

The statistical analysis of recurrent events is additionally complicated by censoring of the event process due to e.g., withdrawal, concomitant medications, ... This censoring process is often informative and may in particular depend on the history of the event process, in which case we say that the censoring process depends on the outcome of interest. For instance, in our motivating application, patients who experience many exacerbations of COPD during the trial are more likely to leave the study. Standard recurrent event analyses then typically give biased results. Indeed, most available methods assume that censoring is solely explained by the baseline covariates in the model. Joint frailty models [DJ08] allow for a dependence between the censoring and recurrent event process that is not explained by baseline covariates. However, their assumption that a fixed (unknown) baseline characteristic can explain this residual dependence, conflicts the typically data-generating process whereby the presence of many events in a given subject makes it more likely to leave the study. Alternatively, it may be tempting to consider additional adjustments for a time-varying covariate that summarizes the history of the recurrent event process at each time. However, such adjustment is inappropriate in studies that aim to learn the effect of a (randomized) treatment on the event process [RHB00]. In this Chapter, we therefore focus on IPCW methods [RR92; RR95]. The use of such techniques

has been considered in the context of Poisson and negative binomial models for recurrent events data [HL99]. We here instead focus on the AG model, as estimation strategies under this model extract more information from the data (by analyzing event times as opposed to numbers of events). IPCW estimators have also been developed under this model [Mil+04], but are rarely used to the best of our knowledge. The aim of this work is therefore to provide a user-friendly introduction to IPCW methods for the AG model.

We conclude this introduction with an overview of the work. In Section 6.2 we present the trial that motivates the use of the presented techniques. In Section 6.3 we review the literature and give insight into its application. Then, in Section 6.4, we apply IPCW for the AG model on simulated data and give user-friendly explanations on how to handle, in practice, the informative withdrawal problem. We highlight potentially problematic settings that should be carefully avoided. In Section 6.5, we provide an analysis of the motivating trial. Finally, we briefly summarize and discuss the results in Section 6.6. Detailed codes, figures, and tables with full results are reported in Section 6.7.

6.2 Motivating example

Chronic Obstructive Pulmonary Disease (COPD) is characterized by persistent, progressive, and only partially reversible airflow obstruction. Patients affected by COPD usually face events that worsen their respiratory symptoms, called acute exacerbations (AECOPDs). These exacerbations are categorized, depending on their acuteness, as mild (controlled with an increased dosage of regular medications), moderate (treated with standard of care: systemic or oral corticosteroids and/or antibiotics), or severe (when hospitalization is required). The rate of these events can vary significantly between patients. An investigational multicomponent vaccine has been developed to reduce the frequency of moderate and severe AECOPDs associated with Non-Typeable Haemophilus influenzae (NTHi) and Moraxella catarrhalis (Mcat), pathogens that are frequently identified in association with AECOPD. The NTHi-Mcat vaccine safety was established in a previous phase 1 study of adults with a smoking history and in a multicentre, randomized, observer-blinded, placebo-controlled, proof-of-concept, phase 2b trial of adults with a history of AECOPD [And+22]. Subjects were assigned in a 1:1 ratio through utilization of a minimization algorithm for the administration of two intramuscular injections of the NTHi–Mcat vaccine or a placebo, spaced 60 days apart, in conjunction with standard care. The allocation algorithm factored in age category, number of prior exacerbations, COPD severity at study initiation, and country, serving as minimization variables to ensure equitable distribution of treatment within each variable. Both vaccine recipients and individuals responsible for evaluating study endpoints were kept unaware of group allocation. In the efficacy analysis, the primary outcome was the rate of any moderate or severe AECOPD occurring within a 1-year period (starting 1 month after the second dose in patients who received two vaccine doses). Safety was assessed in the total vaccinated cohort. The completed trial is registered at `ClinicalTrials.gov` with the number NCT03281876. The vaccine was immunogenic if administered in a two-dose schedule and no safety concerns were identified. In this phase 2b trial, eligible patients aged 40-80 years suffered from diagnosed COPD, with at least one documented moderate or severe event in the previous year. Patients were selected from eight countries: Belgium, Canada, France, Germany, Italy, Spain, the UK, and the USA. They were categorized

by looking at the degree of airflow limitation, according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD grade 2, 3, or 4). They were administered two intramuscular injection doses of the investigational vaccine or placebo and were monitored twice a day, in the morning and evening, using an electronic self-compiled diary. Symptoms of COPD are divided into major (such as dyspnoea, sputum volume, and sputum purulence) and minor ones (such as sore throat, cold, fever, increased cough, and increased wheeze). Whenever at least two major symptoms (or one major symptom and some minor ones) got worse for at least two days, a potential event was recorded by the diary and needed to be confirmed and classified by the investigators. At each medical examination, sputum samples were collected and analyzed. Patients received the first injection on day one, the second injection on day 60, and the efficacy starting date was set to be day 90. The desired follow-up period was 1 year from the efficacy starting date, but some patients withdrew from the study before the end of follow-up. The primary endpoint was the reduction in the yearly rate of moderate and severe events.

In [And+22], a negative binomial model was used to estimate vaccine efficacy in reducing moderate and severe events. The NTHi–Mcat vaccine administered to patients with COPD did not show efficacy in reducing the yearly rate of moderate or severe exacerbations (mean vaccine effect was estimated to be -2.26%). No correction for informative censoring was implemented. Moreover, the analysis highlights the low amount of Severe AECOPD which was interpreted as a potential signal to be further evaluated. For this reason, additional analyses were performed to study the vaccine efficacy w.r.t. reduction of severe exacerbations only, namely only severe events were considered to be events of interest [Aro+22]; vaccine efficacy was then estimated to be 36.54% (CI [4.69%, 61.54%], p 0.08). To understand whether this result is potentially affected by informative censoring, we perform a re-analysis, accounting for the withdrawal mechanism.

To estimate the withdrawal rate we use generalized additive models (GAMs) and Cox models. Both GAMs and Cox models describe the censoring mechanism depending on covariates and a non-parametric baseline function of time. Baseline covariates are age, sex, race, GOLD grade called GOLDGRD, the history of events before entering the study called HISTEXA (< 2 or ≥ 2), country (categorized as Europe or America), and smoker status (current or previous).

The analyzed dataset is composed of 571 protocol-compliant patients, 279 under treatment, and 292 under placebo. A total of 46 patients withdrew from the study, of whom 13 were in the treatment arm and 33 in the placebo arm [And+22]. Figure 6.1 shows the Kaplan-Meier estimator for withdrawal by treatment arm. Kaplan-Meier curves displaying associations with other baseline covariates can be found in Sub-section 6.7.2. While withdrawals are quite balanced with respect to most of the covariates, it is interesting to note the imbalance with respect to treatment (log-rank-test with pvalue 0.001). This may potentially indicate the effectiveness of the treatment received.

We subsequently study the relationship between withdrawal and the outcome itself, to understand if patients with more events are more likely to withdraw. For this, several functions of the cumulative number of events (linear, logarithm, square root, and a categorization of the cumulative number of exacerbations (1, 2, 3, or greater) are added to the censoring models. Comparisons by means of the Akaike information criterion (AIC) [Aka92] and ANOVA tests strongly support the hypothesis that treatment and the cumulative number of exacerbations (pvalues 0.0018 and 0.0004, respectively) are associated with the withdrawal process (see Sub-section 6.7.2 for

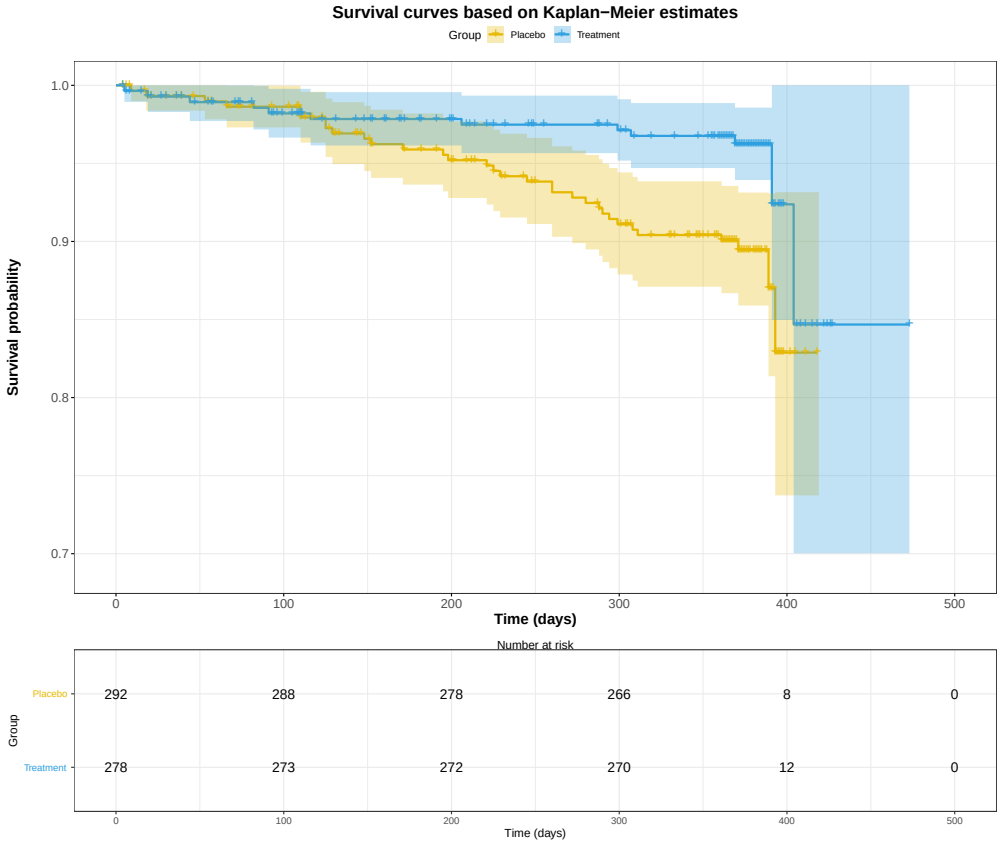

FIGURE 6.1: Kaplan-Meier curves to compare withdrawal probability in treatment (light blue) and placebo (yellow) arms.

further details): in fact, the percentage of withdrawal in treated is around 4.32% and in placebo is around 11.30% (see Figure 6.1). For further details about the design of the trial, the analysis performed, and the results obtained, we refer to the published studies [And+22; Aro+22].

6.3 Methods and Models

The analysis of the considered trial are based on the methods proposed by Miloslavsky et al. [Mil+04]. In particular, consider a randomized clinical trial composed of n patients that received two vaccine doses at day d_1 and day d_2, respectively. Patients are followed from day d_e (such that $d_e > d_2 > d_1$), which is the efficacy starting date, until the end of follow-up τ, which is the same for all patients. The aim of the vaccine is to reduce the number of crises/events that each patient will experience. Vaccine efficacy is be defined in terms of a contrast of event rates in the two arms.

Suppose that for each patient in the study, the following data are collected: the vector of event times $T_i = (d_e \leq T_{i,1}, T_{i,2}, \ldots, T_{i,n_i} \leq \tau)$ and a covariate process $Z_i(\tau)$ (e.g., concomitant medication, time-dependent biomarkers), where n_i is the total number of events that the i-th patient faces during the study and T_{ij} is the time of the j-th event. The covariate process is observed at times t ($d_1 < t < \tau$) which do not need to overlap with event times. We define the recurrent event process which counts the number of events over time to be $N_i(t) = \sum_{k=1}^{n_i} I(T_{i,k} \leq t)$ for each

patient i. Finally, we set $X_i(t) = \{N_i(t), Z_i(t)\}$ to be the full multivariate process of recurrent and covariate events for patient i until time t. $\overline{X}(t) = \{\overline{N}(t), \overline{Z}(t)\}$ is the full multivariate process for the whole population involved in the trial, up to time t. The interest lies in modeling the recurrent event process in function of covariates. For this, we define the intensity of the full recurrent event process $N(t)$ as

$$\begin{aligned} \mathbb{E}(dN(t)|X(t^-)) &= Y_\lambda(t)\lambda(t|X(t^-))dt, \\ \text{with } dN(t) &= N(t+dt) - N(t), \end{aligned} \tag{6.1}$$

where Y_λ is an at-risk indicator defined by the data until time t^-, namely $X(t^-)$, and $\lambda(t|X(t^-))$ is the instantaneous probability of the recurrent events process t jumping at time t conditional on data $X(t^-)$. This intensity can be parameterized using the AG multiplicative intensity model [AG82] as follows:

$$\mathbb{E}(dN(t)|X(t^-)) = Y_\lambda(t)\lambda_{0,f}(t)\exp(\alpha_{f,\lambda}K(t^-))dt, \tag{6.2}$$

with $\lambda_{0,f}(t)$ the baseline intensity function at time t, $\alpha_{f,\lambda}$ the vector of unknown regression coefficients for the full model and $K(t^-)$ a known function of the full data process $\overline{X}(t^-)$. We can think, for example, the covariate process $Z(t)$ to include some baseline individual characteristics (treatment, sex, age, ...) and some time-varying individual measured quantities (concomitant medications prescribed, biomarkers changing over time, ...).
However, often only a subset of the entire covariate process $\overline{Z}(t)$ can be observed; we define the observed covariate process to be $\overline{Z}^*(t)$, such that $\overline{Z}^*(t) \subset \overline{Z}(t)$. The observed process is $\overline{W}(t) = \{\overline{N}(t), \overline{Z^*}(t)\}$, is part of the full data process. We can model this second scenario by looking at the intensity, and conditioning on the observed process

$$\mathbb{E}(dN(t)|W(t^-)) = Y_\lambda(t)\lambda(t|W(t^-))dt = Y_\lambda(t)\lambda_{0,p}(t)\exp(\alpha_{p,\lambda}V(t^-))dt, \tag{6.3}$$

with $\lambda_{0,p}(t)$ the baseline intensity function at time t, $\alpha_{p,\lambda}$ the vector of regression coefficients for the partial observed model and $V(t^-)$ a known function of the observed data process $\overline{W}(t^-)$. To recall the previous example, we may consider a realistic scenario where patients may arbitrarily choose to take other medication on top of the one prescribed by clinicians, to reduce their pain, but no track of all these concomitant medications can be measured and collected in $\overline{Z}^*(t)$.
Moreover, the estimation of the parameter $\alpha_{f,\lambda}$ indexing the AG-model is usually complicated by the fact that some patients may leave the study after time d_e, but before the end of the planned follow-up (withdrawal). We then do not observe the full event and covariate process. In particular, for each patient in the study, a vector $T_i = (d_e \leq T_{i1}, T_{i2}, \ldots, T_{in_i}, \tau \wedge C_i, \Delta_i(\tau))$ is collected, where n_i is the total number of events that patient i experiences during the observation period $\tau \wedge C_i$, with C_i censoring time for patient i, and $\Delta_i(t) = I(t > C_i)$. The multivariate observed individual process is then $X_i^*(t \wedge C_i) = \{N_i^*(t), Z_i^*(t \wedge C_i)\}$, where $N_i^*(t) = \sum_k I(T_{i,k} \leq t \wedge C_i)$. The full complete process is $\overline{X}^*(t \wedge C) = \{\overline{N}^*(t), \overline{Z}^*(t \wedge C))\}$ with $C = (C_i, ..., C_n)$ vector of the censoring times. Finally, the censoring individual process is $U_i(t) = I(C_i < t)$, while $\overline{U}(t) = I(C < t)$ is the whole population process. Its conditional hazard can

also be parameterized using the AG-model:

$$\begin{aligned}\mathbb{E}(dN^*(t)|\overline{X}^*(t^- \wedge C), \overline{A}(t^-)) &= Y^*_\lambda(t)\lambda^*(t|\overline{X}^*(t^- \wedge C), \overline{U}(t^-))dt \\ &= Y^*_\lambda(t)\lambda^*_{0,p}(t)\exp(\alpha^*_{p,\lambda}V^*(t^-))dt,\end{aligned} \tag{6.4}$$

with $\lambda^*_{0,p}(t)$ baseline intensity function at time t, $\alpha^*_{p,\lambda}$ vector of regression coefficients for the model and $V^*(t^-)$ a known function of covariate process $\overline{X^*}(t^- \wedge C)$.

To draw inference for the model in Equations (6.3) and(5.5) in the presence of censoring, we assume that the censoring process coarsens the data at random (CAR), as defined in Assumption 1.

Assumption 1. *Let $\lambda_C(t|\overline{X}(\tau))$ be the censoring hazard function at time t given the full data process $X(\tau)$ up to final time τ, defined as*

$$\lambda_C(t|\overline{X}(\tau)) = \mathbb{E}[U(t)|\overline{U}(t^-) = 0, \overline{X}(\tau)], \tag{6.5}$$

then for right-censored data, the assumption can be mathematically stated as

$$\lambda_C(t|\overline{X}(\tau)) = \lambda_C(t|\overline{X}(t)), \tag{6.6}$$

meaning that, given the full data, the censoring event defining the observed data depends only on the observed part of the data.

Note that this assumption allows for censoring to depend on the history of the covariate and event process. In particular, it allows for the number of events (e.g., COPD exacerbations) to influence the censoring mechanism. Under this assumption, Miloslavsky et al. [Mil+04] proposed to adjust for censoring via inverse probability of censoring weighting, extending the implementation from the time to event setting [Wil+18] to the recurrent events one.

We do not go into details of the theoretical explanation which is out of the scope of this work and can be found in [Mil+04]. We instead describe the IPCW technique from a more intuitive and practical perspective. To first clarify the methodology, from a user perspective, we can sum up the procedure in the following steps:

- fit a model for censoring (with the preferred method) incorporating the interesting covariates. If there is enough evidence of informative censoring, proceed with other steps. The model can be fitted using different methodologies (Cox model for time-to-event, logistic model, see Chapter 4), but a careful analysis should be performed. We suggest performing different models, including different subsets of covariates, and selecting the best model with criteria such as AIC; If there is enough evidence of informative censoring (i.e., the effect of one or more covariates is statistically significant), proceed with other steps, otherwise, the AG estimation procedure can be directly applied;

- ensure the dataset is in the time-person format (i.e., given the list of all times in the dataset, the time-person format requires one row for each of these times for each subject). Then, estimate the probabilities $\hat{S}^0_i(t)$ of remaining uncensored at each time point time t, for each subject i;

- estimate the probabilities $\hat{S}^0_i(t)$ of remaining uncensored at each time point time t given the subset of baseline and, in particular, the time-varying covariates chosen by the user and associated with the censoring (in our motivating examples, the previous number of exacerbations is included);

- estimate the probabilities $\hat{S}_i(t)$ of remaining uncensored at each time point time t, for each subject i, given the subset of baseline and, in particular, the time-varying covariates chosen by the user and associated with the censoring (in our motivating examples, the previous number of exacerbations is included);

- compute the IPCW weights $\hat{W}_i(t)$ and the IPCW stabilized weights $\hat{W}_i^S(t)$, for each time t and each patient i, as

$$\hat{W}_i(t) = \frac{1}{\hat{S}_i(t)} \quad \text{and} \quad \hat{W}_i^S(t) = \frac{\hat{S}_i^0(t)}{\hat{S}_i(t)}; \tag{6.7}$$

- estimate vaccine efficacy with AG model with robust sandwich variance estimator in the absence of informative withdrawal, with subjects weighted according to the IPCW methodology.
`R` programming language provides the `coxph` function, in the `Survival` package [The+], which can be used to fit AG model with `cluster` option for robust variance estimator and `weights` for IPCW correction.

To give an intuitive explanation, the weights we evaluate are used to construct a hypothetical population in which each observation counts as many times as its weight. We can imagine a dataset composed of each row of the original one, repeated as many times as its weight. The rationale behind this is to give each observation a different informativeness and to make it count differently on the final estimation of the parameters of interest, according to the censoring scheme. In particular, at a certain time s, all the information collected for time $t > s$ for a patient who was very likely to withdraw the study before s (according to the baseline and time-varying characteristics highlighted by the censoring model) should be considered to be very informative, while, on the other hand, information collected for patients who are not likely to withdrawal the study should count less in the final estimate. That is why weigths are inversely proportional to the survival probabilities. It may happen, for some pathological cases, that all the information used to estimate the parameters rely on a few patients whose weights turn out to be very high, compared with other patient weights. For this reason, before applying AG-IPCW for the final estimate of interest, it is important to carefully look at the min-max-weigths-ratio, defined as the ratio between the higher and the lower weigth in the whole dataset, to understand the impact of the less and most significant observations in the dataset.

6.4 Simulations

We performe a simulation study to evaluate the finite sample performance of the proposed estimator and to evaluate the weaknesses and strengths of the proposed methodology. We analyze different scenarios with the Monte Carlo technique: we simulate $N = 1000$ datasets, each composed of $n = 250$ patients with a 2000-day follow-up period. The underlying simulation mechanism is common to all scenarios, while the involved parameters may change.

To keep the models simple, only baseline independent covariates $Z_i = (A_i, L_i)$ are considered, where $A_i \sim Ber(0.5)$ is the treatment indicator and $L_i \sim \mathcal{N}(-6.75, 0.25)$ is a continuous covariate. For patient i, the time between of event j, conditioned on observed time $t_{i,j-1}$ of event $j-1$, is simulated using an

exponential distribution with intensity

$$\lambda_i(t) = G_i \lambda_{f,0} \exp(\alpha_{f,\lambda} A_i + \beta_{f,\lambda} L_i). \tag{6.8}$$

We exploit the frailty term[1] $G_i \overset{i.i.d.}{\sim} \Gamma(a, 1/a)$ to ensure a correlation between events of the same patients. Event times are simulated so long as their cumulative sum is smaller than the fixed follow-up time.

The function `simrec` in the `simrec` `R` package [Ing+] is used to simulate the recurrent events data from the AG model.

The censoring mechanism used to simulate withdrawal time is described by the time-dependent hazard

$$\lambda_{C,i}(t) = \lambda_{C,0} \exp(\alpha_C A_i + \beta_C L_i + \mu_C f(N_i(t^-))), \tag{6.9}$$

and simulated as in [Aus12]. To uniquely identify each of the simulated scenarios, we have to specify the vector of involved parameters

$$\theta = (\lambda_{f,0},\ \lambda_{C,0},\ \alpha_{f,\lambda},\ \beta_{f,\lambda},\ \alpha_C,\ \beta_C,\ \mu_C,\ f(\cdot),\ a). \tag{6.10}$$

In our simulation we use

$$\theta = (0.03,\ 0.002,\ -0.4,\ 0.5,\ -0.75,\ 0.4,\ \mu_C,\ \arctan(\sqrt{\cdot}),\ a). \tag{6.11}$$

The function f plays an important role: it determines the impact of the number of critical events on the probability of withdrawal (namely the faster the growth of f, the smaller the censoring times). The used function (i.e., $\arctan(\sqrt{\cdot})$) guarantees a slow increase in the impact of the recurrent event history.
Changing the parameters a and μ_C results in different informativeness of withdrawal event and then different amplitudes of the bias associated with the estimate of the treatment effect $\alpha_{f,\lambda}$ (see Table 6.1).

TABLE 6.1: Bias obtained on treatment coefficients estimation for five configurations which differ for μ_C and a. Conf 1: $\mu_C = 3$, $a = 1$; Conf 2: $\mu_C = 2$, $a = 1$; Conf 3: $\mu_C = 1$, $a = 1$; Conf 4: $\mu_C = 3$, $a = 0.1$; Conf 5: $\mu_C = 3$, $a = 0.5$. Methods applied are AG, AG with exact IPCW (AG exact-IPCW), with exact-stabilized IPCW (AG exact-sIPCW), with GLM-fitted IPCW, with GLM-fitted-stabilized IPCW (AG GLM-IPCW), with Cox-fitted IPCW (AG Cox-sIPCW), with Cox-fitted-stabilized IPCW (AG Cox-sIPCW).

Configuration	1	2	3	4	5
Withdrawal (%)	67.988	53.916	35.004	79.369	73.687
AG	0.177	0.119	0.038	0.026	0.113
AG exact-IPCW	0.136	0.020	0.002	0.029	0.080
AG exact-sIPCW	0.075	0.010	0.001	0.006	0.035
AG GLM-IPCW	0.139	0.020	0.001	0.029	0.084
AG GLM-sIPCW	0.076	0.011	0.000	0.005	0.037
AG Cox-IPCW	0.138	0.023	0.002	0.028	0.084
AG Cox-sIPCW	0.078	0.013	0.001	0.005	0.038

[1] The parametrization for the Gamma distribution is with shape and scale parameters.

Further details on the code used to simulate the data and the code to estimate the parameters are reported in Sub-section 6.7.1.

The data are analyzed with the AG model, with and without weigths. Weigths are evaluated with different techniques to better exploit the working mechanism of the AG-IPCW method; in particular, weights are calculated exactly using the known parameters (i.e., the ones used to simulate) and are estimated with both logistic regression and Cox models (see Sub-section 6.7.1). Furthermore, we analyze the differences between weights and stabilized weigths (see [HR20], Technical point 12.2).

Table 6.1 reports the percentage of withdrawal (WD %) and the bias associated with $\hat{\alpha}_{f,C}$ for five configurations with different levels of μ_C and a: it can clearly be seen that higher levels of μ_C and smaller levels of a correspond to stronger censoring mechanism (resulting in higher percentage of withdrawal). To better analyze the effectiveness of the corrections tested, Table 6.1 also reports the percentage of bias that AG-IPCW removes relative to standard AG, defined as

$$Correction_m(\%) = \frac{|bias_m| \times 100}{|bias_{AG}|} I(|bias_{AG}| > |bias_m|). \tag{6.12}$$

We comment on the results associated with all the configurations in Table 6.2. Detailed graphics can be found in Sub-section 6.7.1. In particular, Figure 6.3 shows the 95% confidence intervals of the N estimated treatment coefficients: the first boxplot, from the left, shows the estimates that we can obtain if we could observe all the patients until the end of follow-up (i.e., without withdrawal), while the others refer to the scenario with censoring (i.e., with withdrawal). The second boxplot reports the estimates obtained with AG when no corrections to account for informative censoring are applied. From the third to the eighth boxplot, there are different estimates obtained with AG-IPCW methods: we report AG combined with exact weigths computation with and without stabilization (AG exact-sIPCW and AG exact-IPCW), AG combined with logistic weights computation with and without stabilization (AG GLM-sIPCW and AG GLM-IPCW), and finally, AG combined with Cox weigths computation with and without stabilization (AG Cox-sIPCW and AG Cox-IPCW). All the details about implementation and methods can be found in Sub-section 6.7.1.

It can be seen that stabilized weigths deliver better performance than simple weights, as they manage to get rid of a good percentage of the bias obtained when no corrections are applied. Moreover, estimating the weights, instead of using exact ones, does not expand the confidence interval and gives corrections quite similar to those based on exact computations. Unsurprisingly, some finite-sample bias remains in all scenarios; results for Configuration 1-5 at larger sample size ($n = 500, 750$), presented in Table 6.3, indeed confirm that the finite-samples bias shrinks with sample size and the coverage probability improves.

We caution the reader that poor results may be obtained under more extreme scenarios (e.g., configurations 1 and 4) as a result of so-called near-positivity violations. As shown by simulation, the magnitude of the weights plays an important role in the final performance of the method: a strong difference between minimum and maximum weight means that there is an observation in the dataset that will be most responsible for the final estimated value. Thus, even if, from a theoretical point of view, correction is guaranteed, numerical problems may arise. Details of weights distributions can be seen in Figure 6.4 (color code is the same as Figure 6.3).

TABLE 6.2: Result of 1000 simulations obtained on treatment coefficients estimation for five configurations which differ for μ_C and a. Configuration 1: $\mu_C = 3$, $a = 1$; Configuration 2: $\mu_C = 2$, $a = 1$; Configuration 3: $\mu_C = 1$, $a = 1$; Configuration 4: $\mu_C = 3$, $a = 0.1$; Configuration 5: $\mu_C = 3$, $a = 0.5$. SD stands for standard deviation, WD for withdrawal, SE for standard error, and COV for coverage. Methods applied are AG, AG with exact IPCW (AG exact-IPCW), with exact-stabilized IPCW (AG exact-sIPCW), with GLM-fitted IPCW, with GLM-fitted-stabilized IPCW (AG GLM-IPCW), with Cox-fitted IPCW (AG Cox-sIPCW), with Cox-fitted-stabilized IPCW (AG Cox-sIPCW).

Configuration 1	Mean	SD	MSE	Cov (%)	Bias	Corrected bias (%)
AG	-0.223	0.156	0.056	53.6	0.177	0.0
AG exact-IPCW	-0.264	0.302	0.110	43.0	0.136	22.9
AG exact-sIPCW	-0.325	0.250	0.068	55.3	0.075	57.7
AG GLM-IPCW	-0.261	0.293	0.105	44.0	0.139	21.5
AG GLM-sIPCW	-0.324	0.242	0.064	54.8	0.076	56.7
AG Cox-IPCW	-0.262	0.287	0.101	42.0	0.138	21.9
AG Cox-sIPCW	-0.322	0.239	0.063	54.6	0.078	56.1
Configuration 2	Mean	SD	MSE	Cov (%)	Bias	Corrected bias (%)
AG	-0.281	0.155	0.038	62.4	0.119	0.0
AG exact-IPCW	-0.380	0.206	0.043	61.0	0.020	83.6
AG exact-sIPCW	-0.390	0.188	0.035	69.0	0.010	91.7
AG GLM-IPCW	-0.380	0.207	0.043	61.2	0.020	83.0
AG GLM-sIPCW	-0.389	0.187	0.035	68.9	0.011	90.7
AG Cox-IPCW	-0.370	0.205	0.042	62.1	0.023	81.1
AG Cox-sIPCW	-0.387	0.187	0.035	68.9	0.013	89.3
Configuration 3	Mean	SD	MSE	Cov (%)	Bias	Corrected bias (%)
AG	-0.362	0.163	0.028	73.9	0.038	0.0
AG exact-IPCW	-0.398	0.171	0.029	74.4	0.002	94.9
AG exact-sIPCW	-0.399	0.167	0.028	76.8	0.001	98.0
AG GLM-IPCW	-0.399	0.170	0.029	74.1	0.001	97.4
AG GLM-sIPCW	-0.400	0.167	0.028	76.7	0.000	99.5
AG Cox-IPCW	-0.398	0.170	0.029	73.7	0.002	95.2
AG Cox-sIPCW	-0.399	0.166	0.028	75.8	0.001	98.3
Configuration 4	Mean	SD	MSE	Cov (%)	Bias	Corrected bias (%)
AG	-0.374	0.115	0.014	92.3	0.026	0.0
AG exact-IPCW	-0.371	0.205	0.043	69.6	0.029	0.0
AG exact-sIPCW	-0.394	0.146	0.021	88.7	0.006	76.4
AG GLM-IPCW	-0.371	0.211	0.045	71.9	0.029	0.0
AG GLM-sIPCW	-0.395	0.147	0.022	89.4	0.005	81.2
AG Cox-IPCW	-0.372	0.204	0.042	69.7	0.028	0.0
AG Cox-sIPCW	-0.395	0.145	0.021	88.1	0.005	80.1
Configuration 5	Mean	SD	MSE	Cov (%)	Bias	Corrected bias (%)
AG	-0.287	0.136	0.031	75.7	0.113	0.0
AG exact-IPCW	-0.320	0.246	0.067	53.2	0.080	29.2
AG exact-sIPCW	-0.365	0.192	0.038	70.7	0.035	69.0
AG GLM-IPCW	-0.316	0.246	0.068	55.3	0.084	25.6
AG GLM-sIPCW	-0.363	0.190	0.037	71.4	0.037	67.3
AG Cox-IPCW	-0.316	0.240	0.065	54.0	0.084	25.9
AG Cox-sIPCW	-0.362	0.189	0.037	71.1	0.038	66.2

TABLE 6.3: Result of 1000 simulations obtained on treatment coefficients estimation for five configurations which differ for μ_C and a. Configuration 1: $\mu_C = 3$, $a = 1$; Configuration 2: $\mu_C = 2$, $a = 1$; Configuration 3: $\mu_C = 1$, $a = 1$; Configuration 4: $\mu_C = 3$, $a = 0.1$; Configuration 5: $\mu_C = 3$, $a = 0.5$. Each configuration was studied for $n = 250, 500, 750$. COV stands for coverage. Methods applied are AG, AG with exact IPCW (AG exact-IPCW), with exact-stabilized IPCW (AG exact-sIPCW), with GLM-fitted IPCW, with GLM-fitted-stabilized IPCW (AG GLM-IPCW), with Cox-fitted IPCW (AG Cox-sIPCW), with Cox-fitted-stabilized IPCW (AG Cox-sIPCW).

	Bias			Cov (%)		
Configuration 1	$n = 250$	$n = 500$	$n = 750$	$n = 250$	$n = 500$	$n = 750$
AG	0.177	0.186	0.183	53.6	77.7	43.7
AG exact-IPCW	0.136	0.138	0.106	43.0	80.3	57.9
AG exact-sIPCW	0.075	0.079	0.106	55.3	88.8	78.0
AG GLM-IPCW	0.139	0.142	0.113	44.0	81.4	59.9
AG GLM-sIPCW	0.076	0.081	0.059	54.8	88.7	78.1
AG Cox-IPCW	0.138	0.140	0.110	42.0	81.2	60.9
AG Cox-sIPCW	0.078	0.082	0.059	54.6	88.8	78.4
Configuration 2	$n = 250$	$n = 500$	$n = 750$	$n = 250$	$n = 500$	$n = 750$
AG	0.119	0.123	0.120	62.4	77.7	72.9
AG exact-IPCW	0.020	0.010	0.002	61.2	80.3	92.2
AG exact-sIPCW	0.010	0.005	0.002	68.9	88.8	95.8
AG GLM-IPCW	0.020	0.010	0.000	62.1	81.4	92.8
AG GLM-sIPCW	0.011	0.005	-0.001	68.9	88.7	96.0
AG Cox-IPCW	0.023	0.013	0.003	61.0	81.2	92.0
AG Cox-sIPCW	0.013	0.007	0.000	69.0	88.8	95.9
Configuration 3	$n = 250$	$n = 500$	$n = 750$	$n = 250$	$n = 500$	$n = 750$
AG	0.038	0.042	0.035	73.9	90.2	94.3
AG exact-IPCW	0.002	0.004	-0.005	74.4	92.0	97.7
AG exact-sIPCW	0.001	0.004	-0.005	76.8	92.4	97.8
AG GLM-IPCW	0.001	0.003	-0.005	74.1	91.9	97.7
AG GLM-sIPCW	0.000	0.003	-0.005	76.7	92.5	97.9
AG Cox-IPCW	0.002	0.004	-0.005	73.7	91.9	97.8
AG Cox-sIPCW	0.001	0.003	-0.004	75.8	92.9	97.6
Configuration 4	$n = 250$	$n = 500$	$n = 750$	$n = 250$	$n = 500$	$n = 750$
AG	0.026	0.028	0.031	92.3	93.9	93.2
AG exact-IPCW	0.029	0.033	0.015	69.6	71.9	83.7
AG exact-sIPCW	0.006	0.013	0.015	88.7	90.3	95.3
AG GLM-IPCW	0.029	0.032	0.015	71.9	72.6	84.7
AG GLM-sIPCW	0.005	0.012	0.003	89.4	90.3	95.2
AG Cox-IPCW	0.028	0.033	0.015	69.7	72.4	84.5
AG Cox-sIPCW	0.005	0.013	0.004	88.1	89.0	94.8
Configuration 5	$n = 250$	$n = 500$	$n = 750$	$n = 250$	$n = 500$	$n = 750$
AG	0.113	0.116	0.114	75.7	75.5	69.7
AG exact-IPCW	0.080	0.073	0.047	53.2	66.0	79.4
AG exact-sIPCW	0.035	0.032	0.047	70.7	86.1	90.5
AG GLM-IPCW	0.084	0.077	0.049	55.3	67.7	80.4
AG GLM-sIPCW	0.037	0.033	0.019	71.4	86.4	90.8
AG Cox-IPCW	0.084	0.075	0.048	54.0	67.1	78.5
AG Cox-sIPCW	0.038	0.034	0.019	71.1	85.2	90.9

6.5 A real case study: COPD trial

In this section, we repeat the randomized clinical trial analysis of Arora et al. [Aro+22], but now accounting for outcome-dependent censoring. Only compliant patients, namely those who received both vaccine doses, are considered. Patient events are part of the analyses only if they happened during the efficacy period (after day 90 and before the end of follow-up). Patients with no severe event in the observation period were counted as patients without event. The dataset analyzed is composed of 571 patients, 279 under treatment, and 292 assigned to placebo. Further details can be found in Section 6.2. In [And+22], the authors performed an analysis for the primary endpoint (moderate and severe events) and for the secondary endpoint (only for severe exacerbation) using a negative binomial (NB) model, without accounting for informative withdrawal. We repeat their analysis on the secondary endpoint, but now performing the AG-IPCW analysis on the same dataset, to account for information carried out by informative withdrawals.

The results obtained are reported in Table 6.4: point estimation of vaccine efficacy $\hat{VE}$ and the corresponding confidence intervals are presented for AG, AG GLM-sIPCW, AG Cox-sIPCW. Three different generalized linear models (GLM 1-2-3) and Cox models (Cox 1-2-3) are used to describe the withdrawal mechanism, depending on different subsets of baseline and time-varying covariates (in particular, the previous number of exacerbations). Models GLM 0 and Cox 0 refer to the case in which correction is applied only for baseline covariates that were significantly associated with withdrawal (level 95%), while other models correct for time-varying covariates. A full description of GLM 0-1-2-3 and Cox 0-1-2-3 models, with details on implementation, can be found in Sub-section 6.7.2, in the paragraph *Models to estimate vaccine efficacy: AG and AG-IPCW*. The recurrent event model only includes the treatment covariate, while generalized linear models and Cox models used to compute IPCW are selected with the AIC criterion and include both the main and interaction effects of baseline covariates. Covariates included are treatment, age, country, history of exacerbations, gold grade score, smoke indicator, and sex. The list of the covariates

TABLE 6.4: Vaccine efficacy estimation ($\hat{VE}$) with 95% confidence interval obtained with AG method without IPCW, AG-IPCW methods with four GLM models for censoring, AG-IPCW methods with four Cox models for censoring. Weigths interval contains the central 95% of the estimated values ($I_{\hat{W}}$). The first line reports the estimate with NB model from [Aro+22] analysis.

Method	$\hat{VE}$	$CI_{\hat{VE}}$ (95%)	$I_{\hat{W}}$ (95%)
NB [Aro+22]	0.365	[-0.469, 0.615]	–
AG	0.337	[-0.116 , 0.607]	–
AG-IPCW			
GLM 0	0.363	[-0.075 , 0.622]	[1.00 1.187]
GLM 1	0.353	[-0.097 , 0.619]	[0.947 1.049]
GLM 2	0.316	[-0.154 , 0.595]	[0.902 1.089]
GLM 3	0.350	[-0.099 , 0.615]	[0.954 1.053]
Cox 0	0.359	[-0.081 , 0.619]	[1.00 1.176]
Cox 1	0.349	[-0.105 , 0.617]	[0.955 1.051]
Cox 2	0.330	[-0.129 , 0.603]	[0.920 1.147]
Cox 3	0.351	[-0.101 , 0.617]	[0.921 1.103]

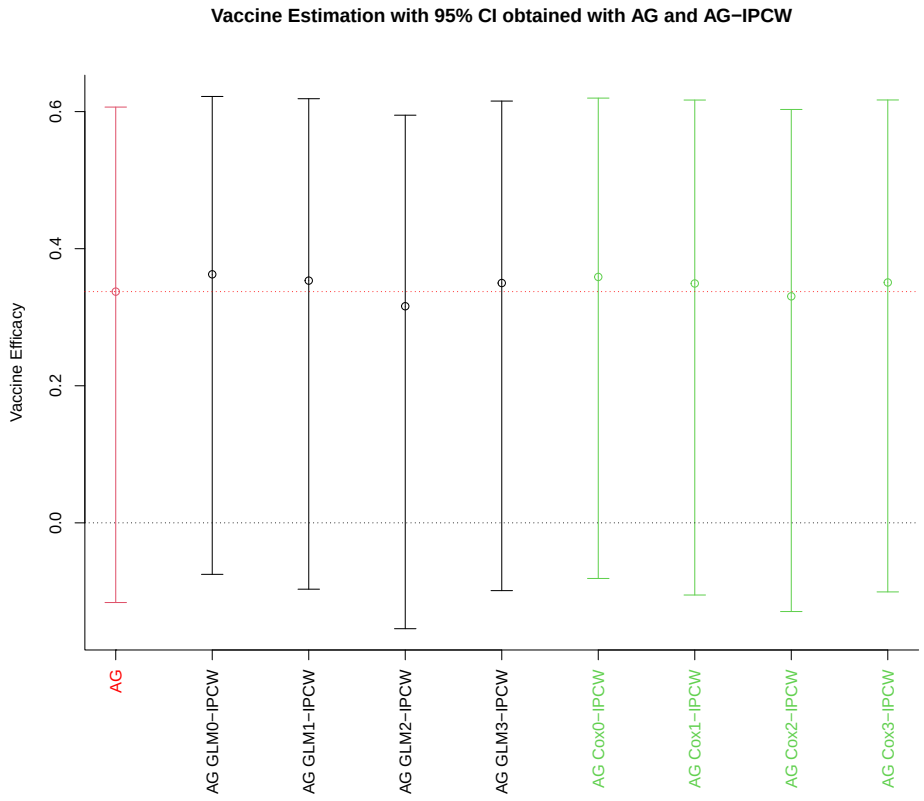

FIGURE 6.2: Vaccine efficacy estimation with 95% confidence interval obtained with AG and AG-IPCW combined with different models for censoring (GLM and Cox). Red dotted line is the mean vaccine efficacy estimated without IPCW correction, the black dotted line is level zero.

included in the model with their full description can be found in Sub-section 6.7.2.

We can reasonably assume, due to the structure of the clinical trial and the information collected by the electronic daily survey, that patients withdraw from the study if their health condition is worsening, which is strictly related both to their previous clinical history (smoking status, history of exacerbations and gold grade), both to the number of exacerbations they experiment. As such, we can think of the censoring mechanism in accordance with the Coarsened at Random (CAR) assumption.

The mean vaccine efficacy estimates, with its 95% confidence interval in Table 6.4, clearly show that IPCW correction does not change the final results, sustaining the result reported in [Aro+22]. This means that informative censoring does not impact the estimation of the treatment effect in the censored data. Even if the withdrawal is significantly different between placebo and treatment, and strongly dependent on the cumulative number of exacerbations, its impact on the treatment effect estimate is not strong. In addition, to give an intuitive visual idea of the results, we report the mean estimated VE with the 95% CI in Figure 6.2. The vaccine estimate, which also accounts for censoring, in accordance with the result published in [Aro+22], suggests a non-significant reduction in adverse events for the vaccinated group.

6.6 Final remarks and conclusions

In this Chapter, we provide a user-friendly guideline on how to apply AG-IPCW to make inference on recurrent event data affected by informative censoring. We explain in details all the necessary steps to perform a comprehensive and accurate analysis of the data, that can be summarized as:

- step (I): study the censoring mechanism with respect to covariates to test the initial hypothesis of informative censoring;
- step (II): in case of a positive result, construct a good model for censoring and use it within the AG-IPCW procedure;
- step (III): carefully examine the distribution of weights.

In Section 6.2 we presented how step (I) should be conducted by mean of the real case study that motivates the work. We performed steps (II) and (III) on both simulations and the real case study (Sections 6.4 and 6.5). Simulations were particularly valuable for testing the AG-IPCW method in various scenarios and highlighting its strengths and weaknesses. The method proved to be quite powerful but showed strong dependence on weight numerical problems. The methodologies used to construct the weights (e.g., GLM and Cox) yielded similar results, demonstrating that the method is robust to the censoring model. However, it is crucial to stabilize the weights to avoid high variations. The simulation study reveales that when the min-max ratio is high, the results are no longer consistent, as the final estimates depend mostly on a few patients with very high weights. Nevertheless, stabilization alone is often insufficient to ensure good results. In the literature, some methods have been studied for inverse probability weighting (IPW) to address numerical issues. One of the most well-known techniques is the covariate balancing IPW, which ensures bounded weights [IR13]. Some works extending it to IPCW are available [YS22], but further studies may represent an interesting field for future research.

In the real case study, the obtained results did not show a significant improvement when compared with standard AG performances. If the censoring mechanism is informative but only results in minimal information loss, then the correction provided by IPCW may be small. Nevertheless, it is good practice to apply IPCW whenever there is evidence of informative censoring from step (I), as the contribution it gives cannot be known in advance. We provide the code used to simulate and analyze the data (steps (I)-(II)-(III)).

The proposed method offers significant advantages over other available methods. Firstly, it maximizes the utilization of available information by relying on the complete observed dataset, encompassing all events and their respective time points. In contrast, some methods only consider the cumulative number of events up to a specific time point. Additionally, while many other methods can address recurrent event problems in the presence of informative censoring caused by baseline covariates, the strength of the proposed method lies in its ability to correct bias arising from time-dependent covariates. In our example, it can handle cumulative numbers of adverse events over time. To illustrate, let's consider the joint frailty model (JFM), which is one of the most renowned models in the literature for managing recurrent event settings with informative censoring. It can be described by the individual recurrent event intensity λ_i and individual hazard of censoring $\lambda_{i,C}$, both

depending on some baseline measured covariates A_i and L_i, as

$$\textbf{RE: } \lambda_i = G_i \lambda_{f,0} \exp(\alpha_{f,\lambda} A_i + \beta_{f,\lambda} L_i), \tag{6.13}$$

$$\textbf{C: } \lambda_{i,C} = G_i^{\alpha} \lambda_{C,0} \exp(\alpha_C A_i + \beta_C L_i), \tag{6.14}$$

with $\theta = \{\lambda_{f,0},\ \lambda_{C,0},\ \alpha_{f,\lambda},\ \beta_{f,\lambda},\ \alpha_C,\ \beta_C\ \alpha\}$ vector of parameters, and $G_i \overset{i.i.d.}{\sim} q(\cdot)$ frailty random effect terms. The JFM model somehow incorporates the informative censoring mechanism by using the shared frailty term G_i, which can capture the assumption that the more adverse events a patient experiences, the higher the probability of study withdrawal. This frailty term establishes a connection between the intensity of recurrent events and the hazard of censoring, with the correlation between the processes becoming stronger as $|\alpha|$ increases. The performance of the JFM is heavily reliant on the frailty distribution $q(\cdot)$. A significant drawback of the JFM is the lack of interpretability of the baseline term G_i on which the model is built. Moreover, it assumes censoring is only driven by time-fixed and not time-varying processes. The AG-IPCW method overcomes this limitation by directly modeling the censoring mechanism based on both baseline and time-dependent covariates. This approach enhances interpretability and, in particular, it provides a more transparent understanding of the censoring process, which is usually driven by things happening during the study and not things occurring at baseline. A comprehensive comparison between AG-IPCW and JFM was beyond the scope of this work but would be an interesting future research direction.

6.7 Code, Plots and Detailed Results

In this section, we present the code used to simulate the dataset and report the full detailed results (plots, tables, `R` outputs, ...) we refer to in the main part of this Chapter. This section is composed of a first part (Sub-section 6.7.1) dedicated to simulations and a second one dedicated to real data analyses (Sub-section 6.7.2).

6.7.1 Code and Simulations

In the following section, we provide our code to simulate a censored dataset, with informative censoring strongly dependent on treatment allocation and the cumulative number of events, that can be then used to test the AG-IPCW methods. To make the code easily usable by other users, we first sum up the structure of the recurrent events (RE) intensity function and the censoring (C) hazard, defined as

$$\textbf{RE: } \lambda_i = G_i \lambda_{F,0} \exp(\alpha_{f,\lambda} A_i + \beta_{f,\lambda} L_i), \tag{6.15}$$

$$\textbf{C: } \lambda_{i,C} = \lambda_{C,0} \exp(\alpha_C A_i + \beta_C L_i + \mu_C f(N_i(t^-))), \tag{6.16}$$

with $\theta = \{\lambda_{f,0},\ \lambda_{C,0},\ \alpha_{f,\lambda},\ \beta_{f,\lambda},\ \alpha_C,\ \beta_C,\ \mu_C,\ f(\cdot)\}$ vector of user-defined parameters, and $G_i \overset{i.i.d.}{\sim} \Gamma(a, 1/a)$ frailty Gamma-distributed random effects.

To clarify the stimulation mechanism of the censoring time from Cox proportional hazards model with time-varying covariates, as explained in Chapter 4, we need to specify the cumulative hazard function H and the inverse H^{-1}. For clarity, in the following formulas, we set $t_{i,j}$ to be the observed time $T_{i,j}$ of event j for patient i (simulated with `simrec` function accordingly to the model (6.15)) and X_i being equal

to $\alpha_C A_i + \beta_C L_i$

$$H_i(t, X_i, N_i(t^-)) = \begin{cases} \lambda_{C,0}\exp(X_i)t, & \text{if } 0 \le t < t_{i,1}, \\ \lambda_{C,0}\exp(X_i)t_{i,1} + \exp(X_i{+}\mu_C f(1))(t{-}t_{i,1}), & \text{if } t_{i,1} \le t < t_{i,2}, \\ \quad\vdots & \\ \sum_{j=1}^{k}\lambda_{C,0}\exp(X_i{+}\mu_C f(j{-}1))(t_{i,j}{-}t_{i,j-1}) & \\ \quad + \exp(X_i{+}\mu_C f(k))(t{-}t_{i,k}), & \text{if } t_{i,k} \le t < t_{i,k+1}, \\ \quad\vdots & \end{cases}$$

and its inverse function H^{-1}

$$H_i^{-1}(t, X_i, N_i(t^-)) = \begin{cases} \dfrac{t}{\lambda_{C,0}\exp(X_i)}, & \text{if } t \in R_1, \\ \dfrac{t{-}\lambda_{C,0}\exp(X_i)t_{i,1}}{\lambda_{C,0}\exp(X_i+\mu_C f(1))} + t_{i,1}, & \text{if } t \in R_2, \\ \quad\vdots & \\ \dfrac{t{-}\sum_{j=1}^{k}\lambda_{C,0}\exp(X_i{+}\mu_C f(j{-}1))(t_{i,j}{-}t_{i,j-1})}{\lambda_{C,0}\exp(X_i{+}\mu_C f(k))} + t_{i,k}, & \text{if } t \in R_{k+1}, \\ \quad\vdots & \end{cases}$$

where

$$\begin{aligned} \mathrm{R}_1 &= [0,\ \lambda_{C,0}\exp(X_i)t_{i,1}], \\ \mathrm{R}_2 &= [\lambda_{C,0}\exp(X_i)t_{i,1},\ \lambda_{C,0}(\exp(X_i)t_{i,1} + \exp(X_i)(t_{i,2} - t_{i,1}))], \\ &\vdots \\ \mathrm{R}_{k+1} &= [\sum_{j=1}^{K}\lambda_{C,0}\exp(X_i + \mu_C f(j-1))(t_{i,j} - t_{i,j-1})), \\ &\qquad \sum_{j=1}^{K+1}\lambda_{C,0}\exp(X_i + \mu_C f(j-1))(t_{i,j} - t_{i,j-1}))], \\ &\vdots \end{aligned} \tag{6.17}$$

The following code implements the functions to evaluate H, H^{-1} and its intervals of definition R_k and samples the censoring time as $H^{-1}(-\log u, X_i, N_i)$, with $u \sim \mathcal{U}(0,1)$.

```
## FUNCTION TO CALCULATE INTERVALS INVERSE FUNCTION
## OF CUMULATIVE HAZARD (Hinv)

# time=c(t1,...,tn,FUP) times of events {ti} and follow-up time (FUP)
# X                     covariates
# theta                 coefficients of X in censoring hazard
# gamma                 coefficient of number of events in censoring
#                       hazard
# lo                    lambda_C,0 baseline of censoring hazard
# u                     sampled uniform
# nexa                  total number of events
# fun_cum               function of the number of event in censoring
#                       hazard
```

```
Rinterval=function(time,X,theta,gamma,lo,u,nexa,fun_cum)
{
  # construct time interval of definition for H
  # check in which interval is the value -log(u)
  # matrix with values on which is defined H
  R=array(dim=c(2,nexa+1))

  indexR=1
  R[1,1]=0
  R[2,1]=lo*exp(theta%*%X)*time[1]*exp(gamma*fun_cum(0))
  for(j in 2:(nexa+1))
  {
    R[1,j]=R[2,j-1]
    R[2,j]=R[1,j]+
           lo*exp(theta%*%X)*(time[j]-time[j-1])*exp(gamma*fun_cum(j-1)
    )
    if(-log(u)>R[1,j] & -log(u)<R[2,j])
    {
      indexR=j
      break
    }
  }

  return(indexR)
}

## FUNCTION TO CALCULATE INVERSE FUNCTION OF CUMULATIVE HAZARD (Hinv)
# time=c(t1,...,tn,FUP) times of events {ti} and follow-up time (FUP)
# X                     covariates
# theta                 coefficients of X in censoring hazard
# gamma                 coefficient of number of events in censoring
    hazard
# lo                    lambda_C,0 baseline of censoring hazard
# u                     sampled uniform
# J                     interval selected
# fun_cum               function of the number of event in censoring
    hazard
Hinv=function(J,time,X,theta,gamma,lo,u,fun_cum)
{
  # if I'm in jth, then I refer to the interval with j-1 events
  ncrisi=J-1

  y=-log(u)
  if(ncrisi>1)
  {difftime=time[1:ncrisi]-c(0,time[1:(ncrisi-1)])}
  else
  {difftime=time[1:ncrisi]}


  num1=y-lo*exp(theta%*%X)*(exp(gamma*fun_cum(0:(ncrisi-1)))%*%
    difftime)
  den1=lo*exp(theta%*%X+fun_cum(ncrisi)*gamma)
  time2=(num1/den1)+time[ncrisi]

  if(time2<0) ## check to see if the result is feasible
  {message('error -  negative time')}
  return(time2)
}

## SIMULATE CENSORING TIME FOR ONE PATIENT
# hazard rate           h(t)=lo*exp(theta*X+gamma*n.exa)
# time=c(t1,...,tn,FUP) times of events {ti} and follow-up time (FUP)
```

```
# X                       covariates
# theta                   coefficients of X in censoring hazard
# gamma                   coefficient of number of events in censoring
    hazard
# lo                      lambda_C,0 baseline of censoring hazard
# nexa                    total number of events
# fun_cum                 function of the number of event in censoring
    hazard
SimulateTime=function(times,X,theta,gamma,lo,n.exa,fun_cum)
{
  u=runif(1)
  if(n.exa>0)
  {
    # calculate R-interval
    r=Rinterval(times,X,theta,gamma,lo,u,n.exa,fun_cum)

    # evaluate Hinv(-log(u))
    if(r>1)
    {cens=Hinv(r,times,X,theta,gamma,lo,u,fun_cum)}
    else
    {cens=-log(u)/(lo*exp(theta%*%X+gamma*fun_cum(0)))}
  }
  else
  {
    cens=-log(u)/(lo*exp(theta%*%X+gamma*fun_cum(0)))
  }

  return(cens)
}

## FUNCTION TO SIMULATE RECURRENT EVENTS DATA  WITHOUT CENSORING
simulate_Data<-function(N,
                        fu.min,
                        fu.max,
                        cens.prob,
                        dist.x,
                        par.x,
                        beta.x,
                        dist.z,
                        par.z,
                        dist.rec,
                        par.rec)
{
  simdata <- simrec(N,
                    fu.min,
                    fu.max,
                    cens.prob,
                    dist.x,
                    par.x,
                    beta.x,
                    dist.z,
                    par.z,
                    dist.rec,
                    par.rec)
  # add the cumulative number of events to the dataset at ending time
  simdata$csum <- ave(simdata$status, simdata$id, FUN=cumsum)
  return(simdata)
}

## FUNCTION TO SIMULATE THE CENSORED DATA
simulate_Censored_Data<-function(simdata,
                                 thetaexact,
```

```
                                    gammaexact,
                                    loexact,
                                    FUP,
                                    fun_cum)
{
  #generate censor time using the user-defined function
  id=unique(simdata$id)
  t=array(NA,dim=c(length(id),1))
  for(bb in 1:length(id)){
    dd=simdata[simdata$id==id[bb],]
    t[bb]=SimulateTime(times=dd$stop[],
                       X=as.numeric(dd[1, c("x.V1","x.V2")]),
                       theta=thetaexact,
                       gamma=gammaexact,
                       lo=loexact,
                       n.exa=max(dd$csum),
                       fun_cum)
  }
  cens.time=t
  t[t>=FUP]=FUP

  #censoring the data
  bb=1
  dd=simdata[simdata$id==id[bb],]
  dd1 <- survSplit(Surv(start,stop,status) ~ .,data=dd, cut=t[bb])
  dd1=dd1[dd1$stop<=t[bb],]
  dd1$cens=0
  if(t[bb]<FUP) dd1$cens=ifelse(dd1$status==1,0,1)
  simdata.cens=dd1
  for(bb in 2: length(id)){
    dd=simdata[simdata$id==id[bb],]
    dd1 <- survSplit(Surv(start,stop,status) ~ .,data=dd, cut=t[bb])
    dd1=dd1[dd1$stop<=t[bb],]
    dd1$cens=0
    if(t[bb]<FUP) dd1$cens=ifelse(dd1$status==1,0,1)
    simdata.cens=rbind(simdata.cens,dd1)
  }
  simdata.cens$event=simdata.cens$status
  return(list(simdata.cens=simdata.cens,cens.time=cens.time))
}
```

The user-defined functions (lines 1–171) of the code above can be exploited to simulate uncensored and censored data, as reported by 1–67 lines of the code below.

```
## MAIN CODE TO SIMULATE THE DATA
# number of patients in each dataset
N=250
# Minimum and Maximum length of follow-up.
# If fu.min=fu.max, then all individuals have a common follow-up.
fu.min =2000
fu.max =2000
# the censoring provided by the package is not used
cens.prob = 0

# distributions of covariates
dist.x <- c("binomial", "normal")
# parameteres of binomial distribution and gaussian distribution
par.x <- list(0.5, c(-6.75, 1/4))
# coefficients of covariates in RE process
beta.x <- c(-0.4,0.5)
# frailty distribution
```

```
dist.z <- "gamma"
# variance of frailty
par.z <-0.1

# lambda_f,0(t) shape of baseline intensity function
dist.rec <- "weibull"
# if the second parameter is 1 it is a constant baseline hazard
par.rec <- c(exp(-3), 1)

# coefficients of covariates in censoring hazard
thetaexact=c(-0.75,+0.4)
# coefficients of cumulative number of events in CE process
gammaexact=3
# baseline effect in CE process
loexact=exp(-6)

# select the relationship between censoring and cumulative number of
    events:
# h(t)=ho*exp(theta*X+gamma*f(N(t-)))
fun_cum=function(j){return(atan(sqrt(j)))}

# simulate data without censoring
simdata <- simulate_Data(N,
                         fu.min,
                         fu.max,
                         cens.prob,
                         dist.x,
                         par.x,
                         beta.x,
                         dist.z,
                         par.z,
                         dist.rec,
                         par.rec)

# censor data
censoredresult<- simulate_Censored_Data(simdata,
                                        thetaexact,
                                        ammaexact,
                                        loexact,
                                        fu.max,
                                        fun_cum)
simdata.cens=censoredresult$simdata.cens
cens.time=censoredresult$cens.time

# add the comulative number of events to the dataset at starting time
simdata.cens$csum_start <-ave(simdata.cens$event,
                              simdata.cens$id,
                              FUN=function(x){cumsum(c(0,x[-length(x)])
    )})
simdata$csum_start <-ave(simdata$status,
                         simdata$id,
                         FUN=function(x){cumsum(c(0,x[-length(x)]))})
```

Once the dataset for the analysis is simulated, it can be analyzed with the AG and AG-IPCW functions. We report the code to perform the analysis, summarized by the main steps presented in Section 6.3. First of all, before applying the estimation procedure, the dataset needs to be set in the time-person format (lines 10–28), where the interesting times we want to split over are determined by COPD events and withdrawals. Then, the estimation procedure can be applied: model 0 (lines 30–33) implements AG without correction, model 1 and 2 (lines 35–50 and 53–65,

respectively) implement AG with exact weights without stabilization and with it, model 3 and 3a (lines 68–83) implement AG with GLM-fitted-weights without stabilization and with it, and finally model 4 and 4a (lines 86–118) implement AG with Cox-fitted-weights without and with stabilization. Boxplots summing up the results, already discussed in Section 6.3, are shown in Figures 6.3 and 6.4.

```
## FUNCTION TO FIT AG AND AG-IPCW
# dd                dataset
# fun_cum           function of the number of event in censoring hazard
# thetaexact        coefficients of X in censoring hazard
# gammaexact        coefficient of number of events in censoring hazard
# loexact           lambda0 baseline of censoring hazard
FunctionAG<-function(dd,fun_cum,loexact,thetaexact,gammaexact)
{
  # define times to split dataset: when an event or a censoring happen
  time=unique(dd$start[dd$event==1 | dd$cens==1])

  # split data over event-times and order the data first by id,
  # then by time interval
  dataset.long <- survSplit(Surv(start,stop,event) ~ .,data=dd, cut=
    time)
  dataset.long <- dataset.long[order(dataset.long$id,dataset.long$start
    ),]

  # repeat for censoring-times
  dataset.long.cens <- survSplit(Surv(start,stop,cens) ~ .,data=dd, cut
    =time)
  dataset.long.cens <- dataset.long.cens[order(dataset.long.cens$id,
                                               dataset.long.cens$start),]

  # merge the long dataset and adding cumulative number of events
  dataset.long$censored <- dataset.long.cens$cens
  # cumulative exacerbations at ending time
  dataset.long$csum <- ave(dataset.long$event, dataset.long$id, FUN=
    cumsum)
  # cumulative exacerbations at starting time
  dataset.long$csum_start <-ave(dataset.long$event, dataset.long$id,
                                FUN=function(x){cumsum(c(0,x[-length(x)]))
    })

  ## MODEL 0 - Unweighted AG model on long dataset
  MODEL0<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),dataset.
    long, timefix = FALSE)
  coxAG_IPCW_long<-MODEL0$coef[1]

  ## MODEL 1 - AG-IPCW with exact weights
  # hazard censoring rate h on each time interval
  dataset.long$lambda=loexact*exp(thetaexact[1]*dataset.long$x.V1+
                                  thetaexact[2]*dataset.long$x.V2+
                                  gammaexact*fun_cum(dataset.long$csum_
    start))
  # integral of h in [start,end]
  dataset.long$exactS=dataset.long$lambda*(dataset.long$stop-dataset.
    long$start)
  # H=sum of integrals
  dataset.long=dataset.long %>% group_by(id) %>% mutate(cumExactS =
    cumsum(exactS))
  # Surv=exp(-H)
  dataset.long$ExactSurv=exp(-dataset.long$cumExactS)
  # W(t)=1/Surv(t): Unstabilized weights
  dataset.long$exactw=1/exp(-dataset.long$cumExactS)
```

```
MODEL1<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),dataset.
  long, weights = exactw,timefix = FALSE)
coxAG_IPCW_exactw<-MODEL1$coef[1]


## MODEL 2 - AG-IPCW with exact stabilized weights
dataset.long= dataset.long[ dataset.long$stop>dataset.long$start,]
dataset.long=dataset.long[complete.cases(dataset.long),]
dataset.long$startsq=dataset.long$start^2
# numerator model 1
hazards.model <- glm(censored ~ start+startsq+x.V1+x.V2 ,
                     family=binomial(), data=dataset.long)
dataset.long$p.noevent=1-predict(hazards.model, type="response")
dataset.long=dataset.long %>% group_by(id) %>% mutate(surv0 = cumprod
  (p.noevent))
 # Stabilized weights
dataset.long$sw=dataset.long$surv0*dataset.long$exactw
MODEL2<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),dataset.
  long, weights = sw,timefix = FALSE)
coxAG_IPCW_exactsw<-MODEL2$coef[1]


## MODEL 3 - AG-IPCW with logistic model for weigths
hazards.model.den <- glm(censored ~ start+startsq+x.V1+x.V2+
                         fun_cum(dataset.long$csum_start),
                         family=binomial(), data=dataset.long)
dataset.long$p.noevent.den=1-predict(hazards.model.den, type="
  response")
dataset.long=dataset.long%>%group_by(id)%>%mutate(surv1=cumprod(p.
  noevent.den))
 # Stabilized weights
dataset.long$fit.sw=dataset.long$surv0/dataset.long$surv1
 # Unstabilized weights
dataset.long$fit.w=1/dataset.long$surv1
MODEL3<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),dataset.
  long,weights = fit.sw,timefix = FALSE)
coxAG_IPCW_fitsw<-MODEL3$coef[1]

MODEL3a<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),
                                       dataset.long, weights = fit.w,
                                       timefix = FALSE)
coxAG_IPCW_fitw<-MODEL3a$coef[1]


## MODEL 4 - AG-IPCW with Cox model for weigths
cens_mod_den<-coxph(Surv(start,stop,censored)~x.V1+x.V2+
                                              fun_cum(csum_start),
                                              dataset.long,
                                              timefix = FALSE)
cens_mod_num<-coxph(Surv(start,stop,censored)~x.V1+x.V2,
                                              dataset.long,timefix =
  FALSE)

id=unique(dataset.long$id)
dataset.long$fit.sw.cox=NA
dataset.long$fit.w.cox=NA
for( i in id){
  current_data_at_cens_times<-dataset.long[dataset.long$id==i,]
  km=survfit(cens_mod_den,newdata=current_data_at_cens_times,
                           id=id,timefix=FALSE)
  survest <- stepfun(km$time, c(1, km$surv))
  w=survest(current_data_at_cens_times$stop)
```

```
102
103
104     km0=survfit(cens_mod_num,newdata=current_data_at_cens_times,
105                                 id=id,timefix=FALSE)
106     survest <- stepfun(km0$time, c(1, km0$surv))
107     w0=survest(current_data_at_cens_times$stop)
108
109      # Stabilized weights
110     dataset.long$fit.sw.cox[dataset.long$id==i]<-w0/w
111      # Unstabilized weights
112     dataset.long$fit.w.cox[dataset.long$id==i]<-1/w
113   }
114
115   MODEL4<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),dataset.
        long, weights = fit.sw.cox,timefix = FALSE)
116   coxAG_IPCW_fitsw.cox<-MODEL4$coef[1]
117   MODEL4a<-coxph(Surv(start,stop,event)~x.V1+x.V2+cluster(id),dataset.
        long, weights = fit.w.cox,timefix = FALSE)
118   coxAG_IPCW_fitw.cox<-MODEL4a$coef[1]
119
120
121   return(list(coxAG_IPCW_long=coxAG_IPCW_long,
122               coxAG_IPCW_exactw=coxAG_IPCW_exactw,
123               coxAG_IPCW_exactsw=coxAG_IPCW_exactsw,
124               coxAG_IPCW_fitsw=coxAG_IPCW_fitsw,
125               coxAG_IPCW_fitw=coxAG_IPCW_fitw,
126               coxAG_IPCW_fitw.cox=coxAG_IPCW_fitw.cox))
127 }
```

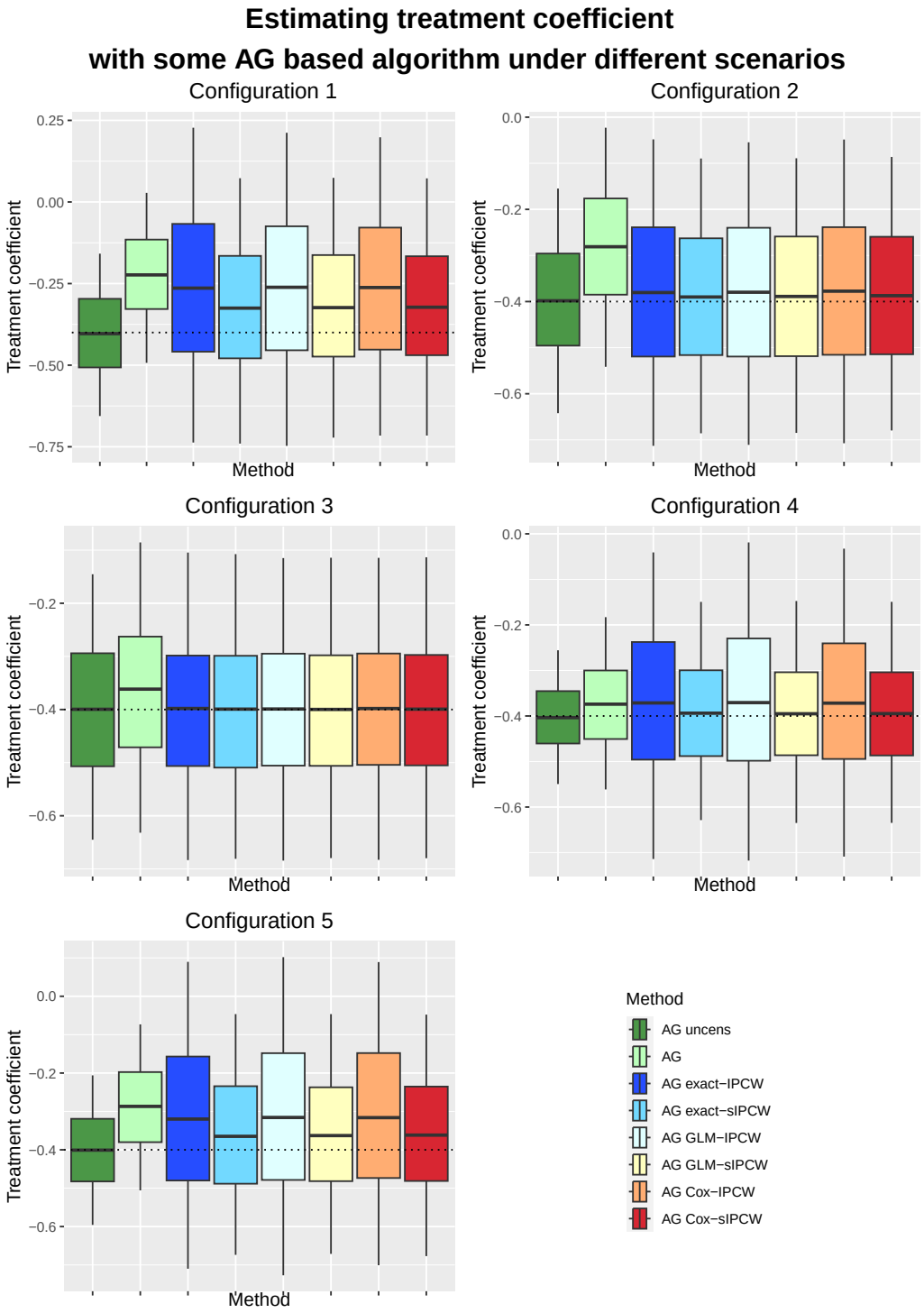

FIGURE 6.3: Boxplot of treatment effect estimates for 1000 datasets obtained with eight estimation procedures. Box shows 25% and 75% quantile, whiskers are 95% and 5%. The solid horizontal line is the mean, while dotted horizontal line is the true vaccine efficacy Methods applied are AG on uncensored data, then AG, AG with exact IPCW (AG exact-IPCW), with exact-stabilized IPCW (AG exact-sIPCW), with GLM-fitted IPCW, with GLM-fitted-stabilized IPCW (AG GLM-IPCW), with Cox-fitted IPCW (AG Cox-sIPCW), with Cox-fitted-stabilized IPCW (AG Cox-sIPCW), on censored data.

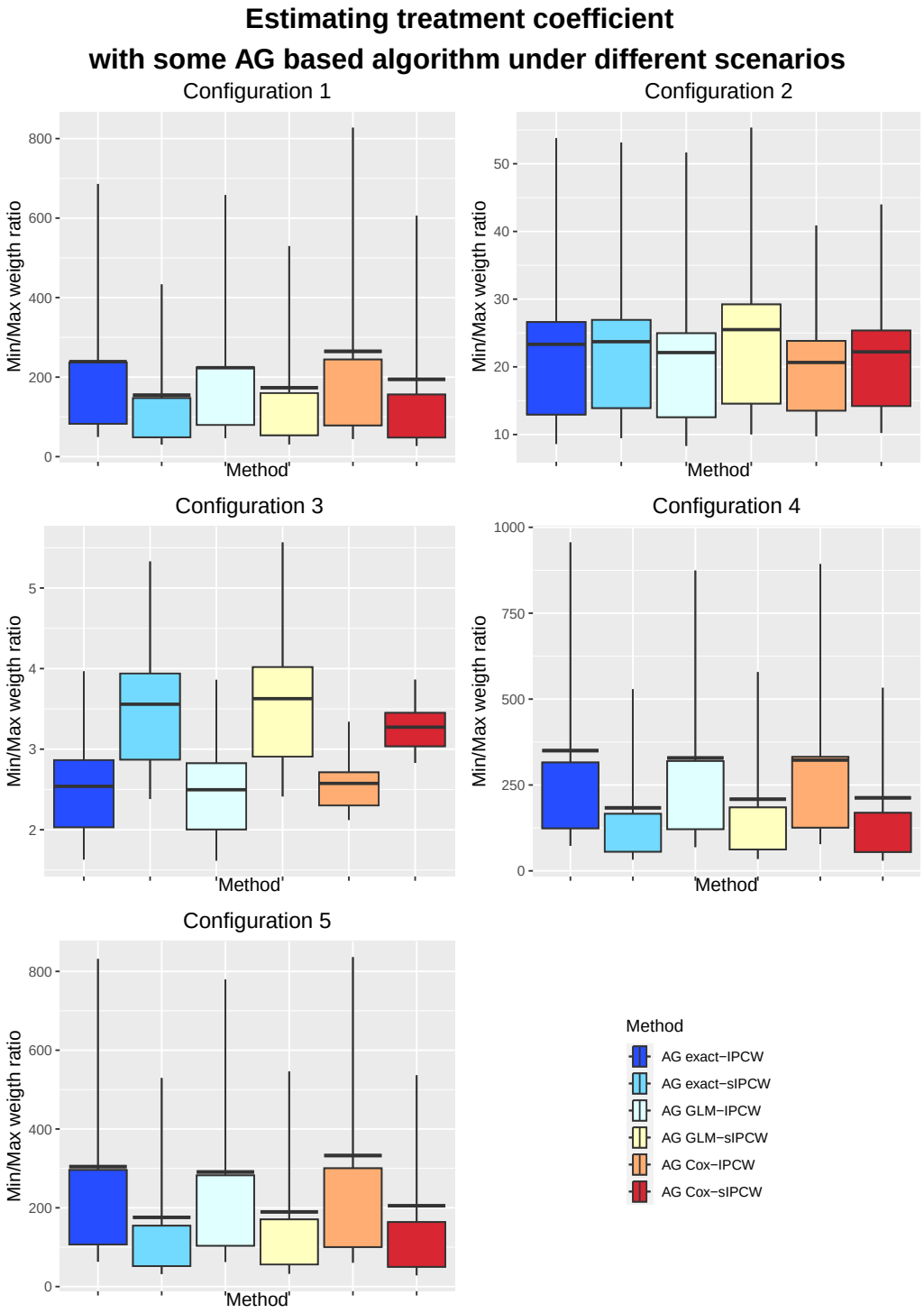

FIGURE 6.4: Boxplot of min-max-weigths-ratio for 1000 datasets obtained with eight estimation procedures. Box shows 25% and 75% quantile, whiskers are 95% and 5%. Solid horizontal line is the mean. Methods applied are AG with exact IPCW (AG exact-IPCW), with exact-stabilized IPCW (AG exact-sIPCW), with GLM-fitted IPCW, with GLM-fitted-stabilized IPCW (AG GLM-IPCW), with Cox-fitted IPCW (AG Cox-sIPCW), with Cox-fitted-stabilized IPCW (AG Cox-sIPCW).

6.7.2 Real data analysis

In this section, we collect the detailed analyses described in the main text about real data provided by GSK. For clarity, we now report and describe in detail the list of baseline covariates that characterized the COPD dataset and that we refer to in the analysis.

- `trt`: dichotomic covariate indicating if the patient is randomized to treatment or placebo;
- `AAGE`: continuous variable indicating exact age;
- `AAGEGR2`: categorical variable indicating the class of age of patients (40-59 or 60-80 years old);
- `ACOUNTRY`: categorical variable indicating the patient nationality (Belgium, Canada, France, Germany, Italy, Spain, United Kingdom, and the United States)
- `country`: dichotomic covariate indicating if the patient is from America or Europe;
- `HISTEXA`: dichotomic covariate indicating if, before enrolment, the patient had more than 2 events or not;
- `GOLDGRD`: degree of airflow limitation. According to the Global Initiative for Chronic Obstructive Lung Disease (levels are 2, 3, or 4).
- `SMOKE`: dichotomic covariate indicating if the patient is currently a smoker or if they were a smoker.
- `SEX`: indicating patient biological sex.

Censoring model

Section 6.2 presents the motivational example and all the analyses performed on the dataset. In particular, the censoring mechanism behind withdrawal is deeply analyzed. We report here the details of the model described and presented in Section 6.2. The ANOVA model used to compare GAMs with different covariates gives the output in Table 6.5 and 6.6, where models tested are:

```
Model 1:  censored ~ s(start) + AGE + HISTEXA + country + GOLDGRD +
                                SEX + SMOKE
Model 2:  censored ~ s(start) + AGE + HISTEXA + country + GOLDGRD +
                           SEX + SMOKE + trt
Model 3:  censored ~ s(start) + AAGE + HISTEXA + country + GOLDGRD +
                         SEX + SMOKE + trt +log.
```

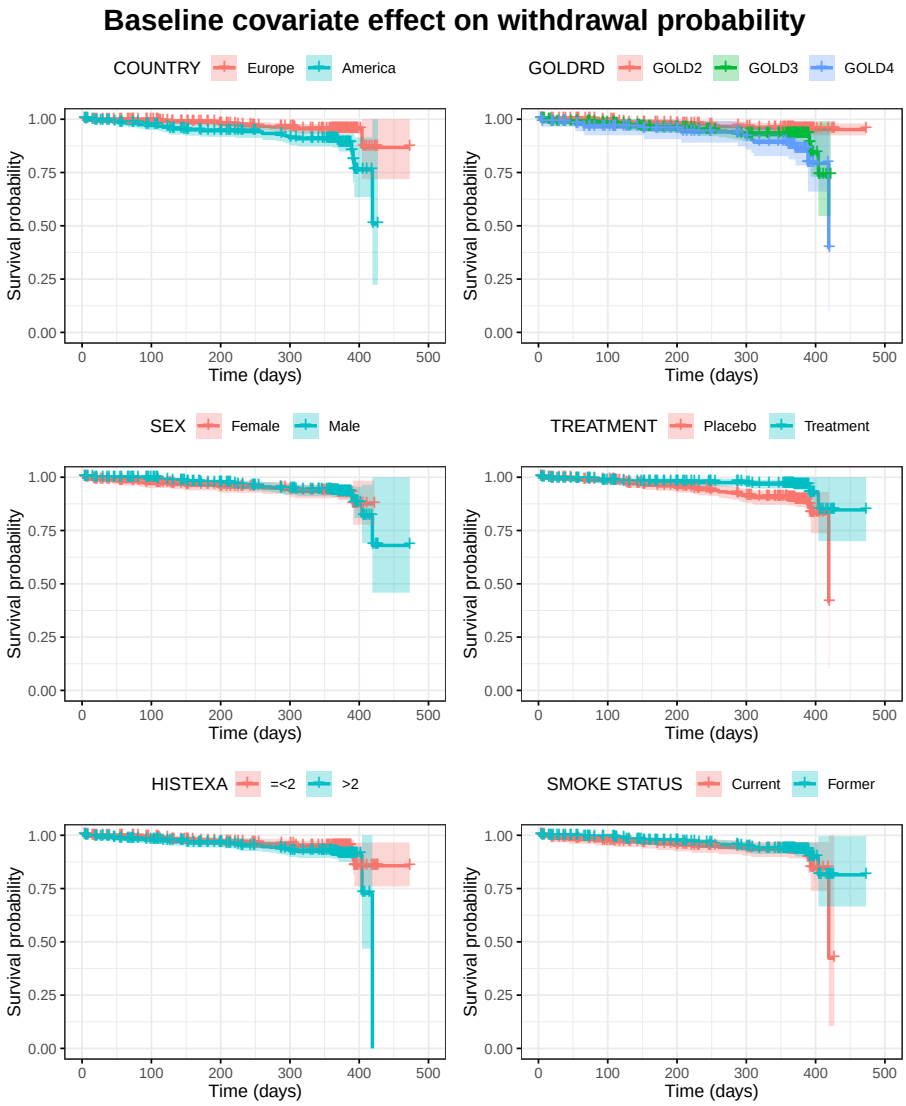

FIGURE 6.5: Kaplan-Meier estimators of survival probability given baseline covariates (COUNTRY, GOLDRD, SEX, TREATMENT, HISTEXA, SMOKE STATUS).

TABLE 6.5: Anova output for GAMs 1 and 2 comparison.

	Resid. Df	Resid. Dev	Df	Deviance	p.value
1	20565	599.01			
2	20564	589.28	1	9.73	0.0018

TABLE 6.6: Anova output for GAMs 2 and 3 comparison.

	Resid. Df	Resid. Dev	Df	Deviance	p.value
2	20564	598.28			
3	20563	589.57	1	12.71	0.0004

To also give a full description of the censoring mechanism we analyze the effect of each baseline covariate in the survival plot in Figure 6.5.
The best Cox model to describe the censoring mechanism, with main effects and interaction terms, was selected using AIC (some covariates were forced to be included in it by the `scope` parameter of the `step` function), by the following code:

```
cox=coxph(Surv(start,stop,censored) ~ (trt + AGE+ HISTEXA + country +
    GOLDGRD + SEX + SMOKE + log)^2 ,data = dataset.long)
bestmodel=step(cox, scope = list(lower = ~ (trt + AGE+ HISTEXA +
    country + GOLDGRD + SEX + SMOKE + trt:log+log)))
```

The best model turned out to have the structure reported in Table 6.7.

TABLE 6.7: Best Cox model output

	term	estimate	std.error	statistic	p.value
1	trt1	-0.31	0.72	-0.43	0.67
2	AGE	0.02	0.05	0.38	0.70
3	HISTEXA>=2	-11.02	3.60	-3.06	0.00
4	country1	1.14	0.38	2.97	0.00
5	GOLDGRDGOLD3	-0.19	0.81	-0.24	0.81
6	GOLDGRDGOLD4	0.19	1.19	0.16	0.88
7	SEXM	4.86	3.40	1.43	0.15
8	SMOKEFormer	-0.46	0.84	-0.55	0.58
9	log	3.18	0.70	4.55	0.00
10	trt1:HISTEXA>=2	1.62	0.83	1.95	0.05
11	trt1:SEXM	-2.46	0.79	-3.13	0.00
12	trt1:log	-1.70	1.03	-1.64	0.10
13	AGE:HISTEXA>=2	0.12	0.05	2.38	0.02
14	AGE:SEXM	-0.10	0.05	-1.93	0.05
15	HISTEXA>=2:GOLDGRDGOLD3	2.22	1.02	2.17	0.03
16	HISTEXA>=2:GOLDGRDGOLD4	3.62	1.24	2.93	0.00
17	HISTEXA>=2:SMOKEFormer	1.36	0.88	1.55	0.12
18	country1:log	-2.49	0.84	-2.96	0.00
19	GOLDGRDGOLD3:SEXM	2.27	0.96	2.37	0.02
20	GOLDGRDGOLD4:SEXM	0.71	0.95	0.74	0.46
21	GOLDGRDGOLD3:SMOKEFormer	-2.45	1.15	-2.13	0.03
22	GOLDGRDGOLD4:SMOKEFormer	-2.38	1.26	-1.89	0.06
23	SEXM:SMOKEFormer	2.08	0.85	2.44	0.01

To evaluate the reliability of the best Cox model selected, we analyze the residual through residual plots in Figures 6.6 and 6.7. The pvalues associated with those residual plots are reported in Table 6.8.

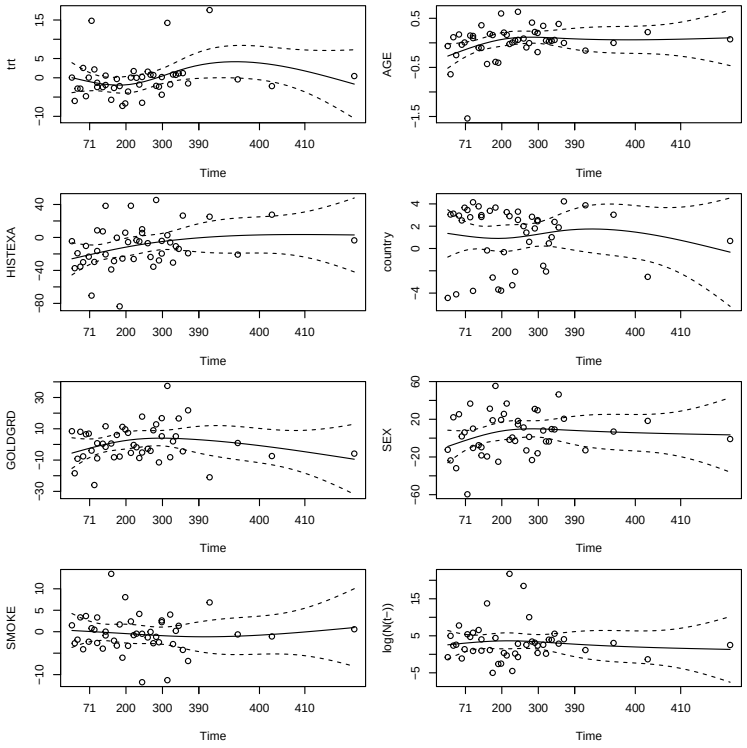

FIGURE 6.6: Schoenfeld's Residual plot for main effects in the Cox model for withdrawal time.

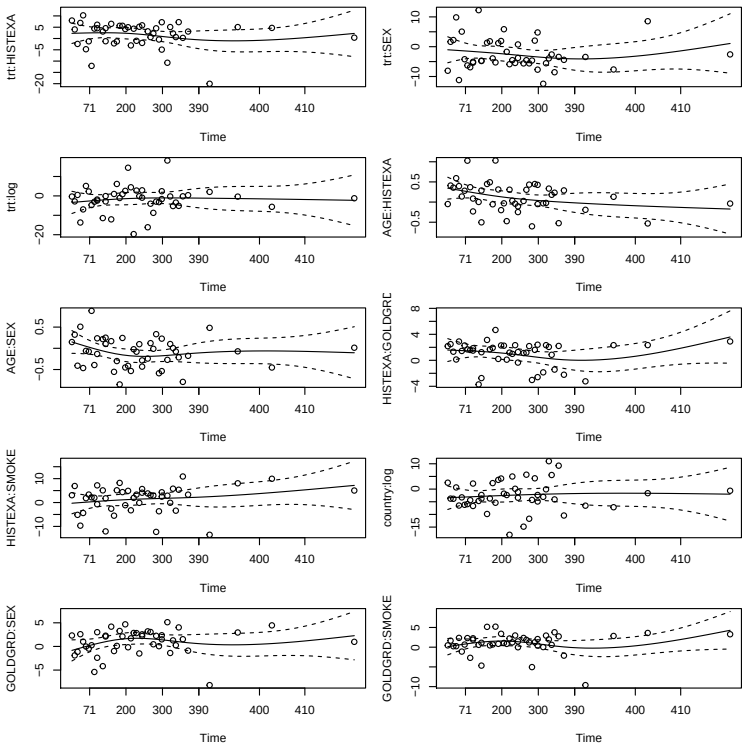

FIGURE 6.7: Schoenfeld's Residual plot for interaction effects in the Cox model for withdrawal time.

TABLE 6.8: Schoenfeld's Residual plot for interaction effects.

term	chisq	df.error	p.value
trt	6.25e-01	1	0.429
AGE	2.01e-01	1	0.654
HISTEXA	1.25e-01	1	0.723
country	4.07e-02	1	0.840
GOLDGRD	2.02e-02	2	0.990
SEX	1.83e+00	1	0.176
SMOKE	5.06e-01	1	0.477
log	8.77e-02	1	0.767
trt:HISTEXA	2.24e-02	1	0.881
trt:SEX	1.37e-01	1	0.711
trt:log	2.70e-03	1	0.959
AGE:HISTEXA	2.11e-03	1	0.963
AGE:SEX	1.26e+00	1	0.262
HISTEXA:GOLDGRD	1.03e-01	2	0.950
HISTEXA:SMOKE	2.03e+00	1	0.154
country:log	6.91e-05	1	0.993
GOLDGRD:SEX	2.79e+00	2	0.248
GOLDGRD:SMOKE	4.92e+00	2	0.085
SEX:SMOKE	2.66e-01	1	0.606
GLOBAL	2.94e+01	23	0.166

Models to estimate vaccine efficacy: AG and AG-IPCW

The code to construct the censoring models (GLM 0-1-2-3 and Cox 0-1-2-3) used to perform the final analyses (see Table 6.4), reported in Section 6.5 are listed here. The models GLM 0-1-2-3 share the structure `glm(formula,family=binomial(),data)`, but differ for the covariates included in the right hand side of the `formula` term. The Cox models 0-1-2-3 share the structure `coxph(formula,data)`, but differ for the covariates included in the right-hand side of the `formula` term. Moreover, each GLM and Cox model is chosen by a search function `step`, based on AIC index.

```
1
2  # create a new dataset without the events associated to level "Spain"
       of ACOUNTRY covariate
3  dataset.long.spain=dataset.long[dataset.long$ACOUNTRY!="Spain",]
4
5  # GLM model 0
6  hazards.model.0 <- glm(censored ~ start + startsq + ( trt + AAGE+
       HISTEXA + country + GOLDGRD +SEX+SMOKE )^2 , family=binomial(),
       data=test.long)
7  best.step.0 <- step(hazards.model.0,scope = list(lower = ~ start+
       startsq+trt+AAGE+country+SEX+SMOKE))
8
9
10 # GLM model 1
11 hazards.model.num <- glm(censored ~ start + startsq + trt + AAGE+
       HISTEXA + ACOUNTRY + GOLDGRD +SEX+SMOKE , family=binomial(), data=
       dataset.long.spain)
12 best.step.num1 <- step(hazards.model.num,scope = list(lower = ~ start+
       startsq+trt+AAGE+ACOUNTRY+SEX+SMOKE))
13
```

```
hazards.model.den <- glm(censored ~ start + startsq + trt + AAGE +
    HISTEXA + ACOUNTRY + log +GOLDGRD+SEX+SMOKE , family=binomial(),
    data=dataset.long.spain)
best.step.den1 <- step(hazards.model.den,scope = list(lower = ~ start+
    startsq+trt+AAGE+ACOUNTRY+log+SEX+SMOKE))

# GLM model 2
hazards.model.num <- glm(censored ~ start + startsq + ( trt + AAGE+
    HISTEXA + ACOUNTRY + GOLDGRD +SEX+SMOKE)^2 , family=binomial(),
    data=dataset.long.spain)
best.step.num2 <- step(hazards.model.num,scope = list(lower = ~ start+
    startsq+trt+AAGE+ACOUNTRY+SEX+SMOKE))

hazards.model.den <- glm(censored ~ start + startsq + ( trt + AAGE+
    HISTEXA + ACOUNTRY + GOLDGRD +SEX+SMOKE+log)^2 , family=binomial(),
     data=dataset.long.spain)
best.step.den2 <- step(hazards.model.den,scope = list(lower = ~ start+
    startsq+trt+AAGE+ACOUNTRY+SEX+SMOKE+log))

# GLM model 3
hazards.model.num <- glm(censored ~ start + startsq + ( trt + AAGE+
    HISTEXA + country + GOLDGRD +SEX+SMOKE )^2 , family=binomial(),
    data=dataset.long)
best.step.num3 <- step(hazards.model.num,scope = list(lower = ~ start+
    startsq+trt+AAGE+country+SEX+SMOKE))

hazards.model.den <- glm(censored ~ start + startsq + (start + startsq
    + trt + AAGE+ HISTEXA + country + GOLDGRD +SEX+SMOKE+log )^2 ,
    family=binomial(), data=dataset.long)
best.step.den3 <- step(hazards.model.den,scope = list(lower = ~ start+
    startsq+trt+AAGE+country+SEX+SMOKE+log))

# Cox 0
cox=coxph(Surv(start,stop,censored) ~(trt + AAGE+ HISTEXA + country +
    GOLDGRD +SEX+SMOKE)^2 ,data = test.long)
  cox0=step(cox)

# Cox numerator
cox=coxph(Surv(start,stop,censored) ~ trt + AAGE+ HISTEXA + ACOUNTRY +
    GOLDGRD +SEX+SMOKE ,data = dataset.long)
cox0=step(cox)

# Cox denominator 1
cox=coxph(Surv(start,stop,censored) ~ trt + AAGE+ HISTEXA + ACOUNTRY +
    GOLDGRD +SEX+SMOKE + log ,data = dataset.long.spain)
cox1=step(cox,scope = list(lower = ~ trt+AAGE+ACOUNTRY+log+SEX+SMOKE))

# Cox denominator 2
cox=coxph(Surv(start,stop,censored) ~ (trt + AAGE+ HISTEXA + ACOUNTRY +
     GOLDGRD +SEX+SMOKE+log)^2,data = dataset.long.spain)
cox2=step(cox,scope = list(lower = ~ trt+AAGE+ACOUNTRY+log+SEX+SMOKE))

# Cox denominator 3
cox=coxph(Surv(start,stop,censored) ~ (trt + AAGE+ HISTEXA + country +
    GOLDGRD +SEX+SMOKE+log)^2 , data = dataset.long)
cox3=step(cox,scope = list(lower = ~ trt+AAGE+country+log+SEX+SMOKE))
```

Chapter 7

Discussion

This thesis comprises three real-data problems that exemplify the challenges encountered when analyzing longitudinal data. It also presents methodologies that can be utilized to model and infer the relevant quantities of interest.

After giving in Chapter 1 some basic knowledge of Bayesian statistic and inference procedures, in Chapter 2 we propose a new compartmental model called SIPRO, which extends the well-known SIR, to analyze the COVID-19 pandemic in Italy, focusing on the estimation of the Not Notified Infected part of the population and its impact on the effective reproductive number. The aim is to construct a realistic model to describe the pandemic mechanism, but simple enough to allow the identifiability of the model parameters. We combine the SIPRO dynamics, in a mixed-effect structure, to describe the heterogeneity of the Italian regions. We infer the model with a Metropolis-within-Gibbs update, combined with Parallel Tempering. We validate the statistical model and the algorithm on simulations and apply it to the data collected by the Italian Protezione Civile during the first phase of the pandemic. We show the improvement gained with our model with respect to the SIR model, with WAIC criterion and comparing short-term predictions. However, we also point out some low-identifiability issues that arise. Future work could include a better study of model identifiability, through methodologies presented in [Cho17; RC19]. Moreover, as new data referring to the pandemic are now available [ISSa], it would be of high interest to first validate the estimates obtained and then incorporate these data into the model to improve its reliability.

In Chapter 3, we analyze data from an observational study involving prostatectomized patients. The study's goal is to identify valuable biomarkers associated with resurgences and enhance the prediction of resurgence timing while optimizing examination timing. To assess tumor recurrence, an expensive PET-PSMA exam is exploited. Given its cost and the importance of minimizing the patient's risk, this exam should be recommended only when strong evidence of tumor recurrence exists. We develop a hierarchical Bayesian model that jointly describes the PSA growth curve and the probability of a positive PET-PSMA; we apply the Gompert model to describe the PSA evolution, accounting for individual patient characteristics with random effects. Since our focus is to estimate the optimal timing for the PET-PSMA exam, we define a statistical procedure to compute the posterior distribution of the optimal time with user-prespecified confidence. The model's performance is evaluated on a simulated dataset and, then, applied to the real dataset. Future work will be devoted to better scheduling the time to collect PSA measurments in order to improve the estimation of the optimal time. Moreover, it would be of high interest to

extend the developed methodology to estimate the time-to-examination in different medical areas.

After a literature review in Chapters 4 and 5, in Chapter 6 we present a study on a recurrent event setting with informative censoring, using the motivating example of COPD provided by GSK. In this context, a new vaccine aims to reduce adverse respiratory events in COPD patients, requiring the estimation of event intensity over time in both the vaccine and placebo arms. However, informative censoring may occur when patients with more events are more likely to withdraw from the study. We demonstrate that parameter estimates in the recurrent event setting can be biased if informative censoring is not taken into account. Our work focuses on analyzing the Inverse Probability of Censoring Weighting (IPCW) technique to address this issue and clarify its practical application in the recurrent event setting. To do this, we combine the IPCW approach with the AG model, building upon the work presented in [Mil+04]. We provide a step-by-step procedure for applying AG-IPCW and highlight potential numerical challenges that may arise, using a simulated dataset. Subsequently, we apply this method to the real dataset provided by GSK. As part of our future work, we aim to further investigate the weight estimation procedure to overcome the numerical challenges encountered in our study, with the ultimate goal of improving the method. Finally, it would be of high interest to make a comparison between the joint frailty model, which is one of the most used techniques in these settings, and the AG-IPCW.

Bibliography

[ABM21] Martina Amongero, Enrico Bibbona, and Gianluca Mastrantonio. "Analysing the Covid-19 pandemic in Italy with the SIPRO model". In: *Book of short papers - SIS 2021* (2021), pp. 1568–1573. URL: https://it.pearson.com//docenti/universita/partnership/sis.html.

[Afs+17] Ali Afshar-Oromieh et al. "Diagnostic performance of ^{68}Ga-PSMA-11 (HBED-CC) PET/CT in patients with recurrent prostate cancer: evaluation in 1007 patients". In: *European Journal of Nuclear Medicine and Molecular Imaging* 44.8 (2017), pp. 1258–1268. DOI: 10.1007/s00259-017-3711-7.

[AG82] Per K. Andersen and Richard D. Gill. "Cox's regression model for counting processes: a large sample study". In: *The Annals of Statistics* (1982), pp. 1100–1120. DOI: 10.1214/aos/1176345976.

[AK15] David F. Anderson and Thomas G. Kurtz. *Stochastic Analysis of Biochemical Systems*. Springer Cham, 2015. DOI: 10.1007/978-3-319-16895-1.

[Aka+18] Mouna Akacha et al. *Request for CHMP qualification opinion: clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses*. Tech. rep. European Medicines Agency, 2018. URL: https://www.ema.europa.eu/en/documents/other/qualification-opinion-treatment-effect-measures-when-using-recurrent-event-endpoints-applicants_en.pdf.

[Aka92] Hirotogu Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Springer New York, 1992, pp. 610–624. DOI: 10.1007/978-1-4612-0919-5_38.

[Ala+21] Pierfrancesco Alaimo Di Loro et al. "Nowcasting COVID-19 incidence indicators during the Italian first outbreak". In: *Statistics in Medicine* 40.16 (2021), pp. 3843–3864. DOI: 10.1002/sim.9004.

[AM06] Christophe Andrieu and Éric Moulines. "On the ergodicity properties of some adaptive MCMC algorithms". In: *The Annals of Applied Probability* 16.3 (2006), pp. 1462–1505. DOI: 10.1214/105051606000000286.

[AM92] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992. URL: https://global.oup.com/academic/product/infectious-diseases-of-humans-9780198540403.

[AMM22] Luis Alvarez, Jean-David Morel, and Jean-Michel Morel. "Modeling COVID-19 Incidence by the Renewal Equation after Removal of Administrative Bias and Noise". In: *Biology* 11.4 (2022), p. 540. DOI: 10.3390/biology11040540.

[Amo+20] Martina Amongero et al. "Development of a Prediction Score to Avoid Confirmatory Testing in Patients With Suspected Primary Aldosteronism". In: *The Journal of Clinical Endocrinology & Metabolism* 106.4 (2020), pp. 1708–1716. DOI: 10.1210/clinem/dgaa974.

[Amo+23] Martina Amongero et al. *Estimating the optimal time to perform a PET-PSMA exam in prostatectomized patients based on data from clinical practice*. 2023. arXiv: 2302.10861.

[And+22] Stefan Andreas et al. "Non-typeable Haemophilus influenzae-Moraxella catarrhalis vaccine for the prevention of exacerbations in chronic obstructive pulmonary disease: a multicentre, randomised, placebo-controlled, observer-blinded, proof-of-concept, phase 2b trial". In: *The Lancet Respiratory Medicine* 10.5 (2022), pp. 435–446. DOI: 10.1016/S2213-2600(21)00502-6.

[Aro+22] Ashwani K. Arora et al. "A detailed analysis of possible efficacy signals of NTHi-Mcat vaccine against severe COPD exacerbations in a previously reported randomised phase 2b trial". In: *Vaccine* 40.41 (2022), pp. 5924–5932. DOI: 10.1016/j.vaccine.2022.08.053.

[AT08] Christophe Andrieu and Johannes Thoms. "A tutorial on adaptive MCMC". In: *Statistics and computing* 18 (2008), pp. 343–373. DOI: 10.1007/s11222-008-9110-y.

[Aus12] Peter C. Austin. "Generating survival times to simulate Cox proportional hazards models with time-varying covariates". In: *Statistics in Medicine* 31.29 (2012), pp. 3946–3958. DOI: 10.1002/sim.5452.

[Aus13] Peter C. Austin. "Correction: Generating survival times to simulate Cox proportional hazards models with time-varying covariates". In: *Statistics in Medicine* 32.6 (2013), pp. 1078–1078. DOI: 10.1002/sim.5723.

[BAB05] Ralf Bender, Thomas Augustin, and Maria Blettner. "Generating survival times to simulate Cox proportional hazards models". In: *Statistics in Medicine* 24.11 (2005), pp. 1713–1723. DOI: 10.1002/sim.2059.

[BCV21] Michela Baccini, Giulia Cereda, and Cecilia Viscardi. "The first wave of the SARS-CoV-2 epidemic in Tuscany (Italy): A SI^2R^2D compartmental model with uncertainty evaluation". In: *PLOS ONE* 16.4 (2021), pp. 1–23. DOI: 10.1371/journal.pone.0250029.

[Ben+08] Luigi Benecchi et al. "Optimal Measure of PSA Kinetics to Identify Prostate Cancer". In: *Urology* 71.3 (2008), pp. 390–394. ISSN: 0090-4295. DOI: https://doi.org/10.1016/j.urology.2007.10.021. URL: https://www.sciencedirect.com/science/article/pii/S0090429507022716.

[BL22] Bruno Buonomo and Deborah Lacitignola. "L'epidemia di COVID-19 in Italia: indagini e risposte dai modelli compartimentali". In: *Matematica, Cultura e Società* 7.1 (2022), pp. 35–52. URL: http://www.bdim.eu/item?id=RUMI_2022_1_7_1_35_0.

[Bon+21] Gianluca Bonifazi et al. "A simplified estimate of the effective reproduction number R_t using its relation with the doubling time and application to Italian COVID-19 data". In: *The European Physical Journal Plus* 136 (2021), p. 386. DOI: 10.1140/epjp/s13360-021-01339-6.

[Bro98] Stephen Brooks. "Markov chain Monte Carlo method and its application". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 69–100. DOI: 10.1111/1467-9884.00117.

[Buf+21] Fabrizio Buffolo et al. "Quality of life in primary aldosteronism: A prospective observational study". In: *European Journal of Clinical Investigation* 51.3 (2021), e13419. DOI: https://doi.org/10.1111/eci.13419.

[Bur+20a] Jacopo Burrello et al. "Prediction of hyperaldosteronism subtypes when adrenal vein sampling is unilaterally successful". In: *European Journal of Endocrinology* 183.6 (2020), pp. 657–667. DOI: 10.1530/EJE-20-0656.

[Bur+20b] Jacopo Burrello et al. "Sphingolipid composition of circulating extracellular vesicles after myocardial ischemia". In: *Scientific Reports* 10 (2020), p. 161820. DOI: 10.1038/s41598-020-73411-7.

[Bur+21] Jacopo Burrello et al. "Characterization of Circulating Extracellular Vesicle Surface Antigens in Patients With Primary Aldosteronism". In: *Hypertension* 78.3 (2021), pp. 726–737. DOI: 10.1161/HYPERTENSIONAHA.121.17136.

[Bur+22a] Jacopo Burrello et al. "Profiling circulating extracellular vesicle surface antigens in primary aldosteronism". In: *Journal of Hypertension* 40.1 (2022), e2. DOI: 0.1097/01.hjh.0000835552.16067.f5.

[Bur+22b] Jacopo Burrello et al. "Supervised and unsupervised learning to define the cardiovascular risk of patients according to an extracellular vesicle molecular signature". In: *Translational Research* 244 (2022), pp. 114–125. DOI: 10.1016/j.trsl.2022.02.005.

[Cal+21] Jamie M. Caldwell et al. "Vaccines and variants: Modelling insights into emerging issues in COVID-19 epidemiology". In: *Paediatric Respiratory Reviews* 39 (2021), pp. 32–39. DOI: 10.1016/j.prrv.2021.07.002.

[Cap93] Vincenzo Capasso. *Mathematical Structures of Epidemic Systems*. Springer Berlin, 1993. DOI: 10.1007/978-3-540-70514-7.

[Car+92] H. Ballentine Carter et al. "Longitudinal evaluation of prostate-specific antigen levels in men with and without prostate disease". In: *JAMA* 267.16 (1992), pp. 2215–2220. DOI: 10.1001/jama.1992.03480160073037.

[Cho17] Gerardo Chowell. "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts". In: *Infectious Disease Modelling* 2.3 (2017), pp. 379–398. DOI: 10.1016/j.idm.2017.08.001.

[CKS19] Anaïs Charles-Nelson, Sandrine Katsahian, and Catherine Schramm. "How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery". In: *Statistics in Medicine* 38.18 (2019), pp. 3476–3502. DOI: 10.1002/sim.8168.

[CL07] Richard J. Cook and Jerald F. Lawless. *The statistical analysis of recurrent events*. Springer New York, 2007. DOI: 10.1007/978-0-387-69810-6.

[Cox+50] Gertrude M. Cox et al. "Discussion on Professor Cox's paper". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 12.1 (1950), pp. 13–18. DOI: 10.1111/j.2517-6161.1950.tb00039.x.

[Cox72] David R. Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–220. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[CP12] Serena H. Chen and Carmel A. Pollino. "Good practice in Bayesian network modelling". In: *Environmental Modelling & Software* 37 (2012), pp. 134–145. DOI: 10.1016/j.envsoft.2012.03.012.

[CVB22] Giulia Cereda, Cecilia Viscardi, and Michela Baccini. "Combining and comparing regional SARS-CoV-2 epidemic dynamics in Italy: Bayesian meta-analysis of compartmental models and global sensitivity analysis". In: *Frontiers in Public Health* 10 (2022). DOI: 10.3389/fpubh.2022.919456.

[Del+20] Fabio Della Rossa et al. "A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic". In: *Nature Communications* 11 (2020), p. 5106. DOI: 10.1038/s41467-020-18827-5.

[Del+23] Giulia Della Croce Di Dojola et al. "Estimating the contagiousness ratio between two viral strains". In: *medRxiv* (2023). DOI: 10.1101/2023.04.27.23289192.

[DJ08] Luc Duchateau and Paul Janssen. *The Frailty Model*. Springer New York, 2008. DOI: 10.1007/978-0-387-72835-3.

[DPCa] Dipartimento della Protezione Civile DPC. *COVID-19 Italia – Monitoraggio situazione*. URL: https://github.com/pcm-dpc/COVID-19.

[DPCb] Dipartimento della Protezione Civile DPC. *COVID-19 Italia – Monitoraggio situazione*. URL: https://github.com/pcm-dpc/COVID-19/blob/master/README_EN.md.

[Ege+02] Lars Egevad et al. "Prognostic value of the Gleason score in prostate cancer". In: *BJU international* 89.6 (2002), pp. 538–542. DOI: 10.1046/j.1464-410X.2002.02669.x.

[Eib+16] Matthias Eiber et al. "^{68}Ga-labeled Prostate-specific Membrane Antigen Positron Emission Tomography for Prostate Cancer Imaging: The New Kid on the Block-Early or Too Early to Draw Conclusions?" In: *European Urology* 70.6 (2016), pp. 938–940. DOI: 10.1016/j.eururo.2016.07.045.

[Fen+19] Wolfgang P. Fendler et al. "Assessment of ^{68}Ga-PSMA-11 PET accuracy in localizing recurrent prostate cancer: a prospective single-arm clinical trial". In: *JAMA oncology* 5.6 (2019), pp. 856–863. DOI: 10.1001/jamaoncol.2019.0096.

[Fos+19] Nicola Fossati et al. "The emerging role of PET-CT scan after radical prostatectomy: still a long way to go". In: *The Lancet Oncology* 20.9 (2019), pp. 1193–1195. DOI: 10.1016/S1470-2045(19)30501-7.

[FP20] Duccio Fanelli and Francesco Piazza. "Analysis and forecast of COVID-19 spreading in China, Italy and France". In: *Chaos, Solitons & Fractals* 134 (2020), p. 109761. DOI: 10.1016/j.chaos.2020.109761.

[Fra07] Christophe Fraser. "Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic". In: *PLOS ONE* 2.8 (Aug. 2007), pp. 1–12. DOI: 10.1371/journal.pone.0000758.

[FRL22] Annalisa Fierro, Silvio Romano, and Antonella Liccardo. "Vaccination and variants: Retrospective model for the evolution of Covid-19 in Italy". In: *PLOS ONE* 17.7 (2022), pp. 1–25. DOI: 10.1371/journal.pone.0265159.

[Gae20] Giuseppe Gaeta. "Social distancing versus early detection and contacts tracing in epidemic management". In: *Chaos, Solitons & Fractals* 140 (2020), p. 110074. DOI: 10.1016/j.chaos.2020.110074.

[Gat+20] Marino Gatto et al. "Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures". In: *Proceedings of the National Academy of Sciences* 117.19 (2020), pp. 10484–10491. DOI: 10.1073/pnas.2004978117.

[GHV14] Andrew Gelman, Jessica Hwang, and Aki Vehtari. "Understanding predictive information criteria for Bayesian models". In: *Statistics and computing* 24 (2014), pp. 997–1016. DOI: 10.1007/s11222-013-9416-2.

[Gio+20] Giulia Giordano et al. "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy". In: *Nature Medicine* 26 (2020), pp. 855–860. DOI: 10.1038/s41591-020-0883-7.

[Gio+21] Giulia Giordano et al. "Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy". In: *Nature Medicine* 27 (2021), 993—998. DOI: 10.1038/s41591-021-01334-5.

[GL06] Dani Gamerman and Hedibert F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Second Edition. Chapman & Hall/CRC, 2006. DOI: 10.1201/9781482296426.

[GM20] Giorgio Guzzetta and Stefano Merler. *Stime della trasmissibilità di SARS-CoV-2 in Italia*. Istituto Superiore di Sanità ISS. 2020. URL: https://www.epicentro.iss.it/coronavirus/open-data/rt.pdf.

[Gna+21] Janyce Eunice Gnanvi et al. "On the reliability of predictions on Covid-19 dynamics: A systematic and critical review of modelling techniques". In: *Infectious Disease Modelling* 6 (2021), pp. 258–272. DOI: 10.1016/j.idm.2020.12.008.

[GRS95] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1995. DOI: https://doi.org/10.1201/b14835.

[Gum+04] Abba B. Gumel et al. "Modelling strategies for controlling SARS outbreaks". In: *Proceedings of the Royal Society B: Biological Sciences* 271 (2004), pp. 2223–2232. DOI: 10.1098/rspb.2004.2800.

[Hir+12] Yoshito Hirata et al. "Quantitative mathematical modeling of PSA dynamics of prostate cancer patients treated with intermittent androgen suppression". In: *Journal of Molecular Cell Biology* 4.3 (2012), pp. 127–132. DOI: 10.1093/jmcb/mjs020.

[HL99] Nicholas J. Horton and Stuart R. Lipsitz. "Review of software to fit generalized estimating equation regression models". In: *The American Statistician* 53.2 (1999), pp. 160–169. DOI: 10.1080/00031305.1999.10474451.

[Hof+19] Manuela A. Hoffmann et al. "The positivity rate of ^{68}Gallium-PSMA-11 ligand PET/CT depends on the serum PSA-value in patients with biochemical recurrence of prostate cancer". In: *Oncotarget* 10.58 (2019), pp. 6124–6137. DOI: 10.18632/oncotarget.27239.

[HR20] Miguel A. Hernan and Jamie M. Robins. *Causal inference: What if*. Chapman & Hall/CRC, 2020. DOI: 10.1201/9781315374932.

[IAR] International Agency for Research on Cancer IARC. *Cancer Today*. [Online]. URL: https://gco.iarc.fr/today/online-analysis-pie.

[Ing+] Katharina Ingel et al. *simrec: Simulation of Recurrent Event Data for Non-Constant Baseline Hazard*. R package. [Online]. URL: https://cran.r-project.org/web/packages/simrec/index.html.

[IR13] Kosuke Imai and Marc Ratkovic. "Covariate Balancing Propensity Score". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2013), pp. 243–263. DOI: 10.1111/rssb.12027.

[ISSa] Istituto Superiore di Sanita ISS. *COVID-19 Italia – Monitoraggio situazione*. [Online]. URL: https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati.

[ISSb] Istituto Superiore di Sanità ISS. *Coronavirus*. [Online]. URL: https://www.epicentro.iss.it/en/coronavirus.

[Ita] Governo Italiano. *Report Vaccini Anti COVID-19*. [Online]. URL: https://www.governo.it/it/cscovid19/report-vaccini.

[Jah08] Antje Jahn-Eimermacher. "Comparison of the Andersen–Gill model with poisson and negative binomial regression on recurrent event data". In: *Computational Statistics & Data Analysis* 52.11 (2008), pp. 4989–4997. DOI: 10.1016/j.csda.2008.04.009.

[JKL19] Alexander Jordan, Fabian Krüger, and Sebastian Lerch. "Evaluating Probabilistic Forecasts with scoringRules". In: *Journal of Statistical Software* 12 (2019), pp. 1–37. DOI: 10.18637/jss.v090.i12.

[Kas+] Alboukadel Kassambara et al. *survminer: Drawing Survival Curves using 'ggplot2'*. R package. URL: https://https://cran.r-project.org/web/packages/survminer/index.html.

[KK12] David G. Kleinbaum and Mitchel Klein. *Survival Analysis - A Self-Learning Text, Third Edition*. New York: Springer, 2012.

[KM27] William Ogilvy Kermack and Anderson Gray McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 115.772 (1927), pp. 700–721. DOI: 10.1098/rspa.1927.0118.

[KM58] Edward L. Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452.

[Kon+22] Lingcai Kong et al. "Compartmental structures used in modeling COVID-19: a scoping review". In: *Infectious Diseases of Poverty* 11 (2022), p. 72. DOI: 10.1186/s40249-022-01001-y.

[KR08] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008. DOI: 10.1515/9781400841035.

[Lav+20] Enrico Lavezzo et al. "Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo". In: *Nature* 584 (2020), pp. 425–429. DOI: 10.1038/s41586-020-2488-1.

[Lee87] Lawrence M. Leemis. "Variate generation for accelerated life and proportional hazards models". In: *Operations Research* 35.6 (1987), pp. 892–894. DOI: 10.1287/opre.35.6.892.

[Lui+20] Henk B. Luiting et al. "Optimal Timing of Prostate Specific Membrane Antigen Positron Emission Tomography/Computerized Tomography for Biochemical Recurrence after Radical Prostatectomy". In: *The Journal of Urology* 204.3 (2020), pp. 503–510. DOI: 10.1097/JU.0000000000001012.

[LUV08] Hans Lilja, David Ulmert, and Andrew J. Vickers. "Prostate-specific antigen and prostate cancer: prediction, detection and monitoring". In: *Nature Reviews Cancer* 8 (2008), pp. 268–278. DOI: 10.1038/nrc2351.

[LZ20] Elena Loli Piccolomini and Fabiana Zama. "Monitoring Italian COVID-19 spread by a forced SEIRD model". In: *PLOS ONE* 15.8 (Aug. 2020), pp. 1–17. DOI: 10.1371/journal.pone.0237417.

[Mar15] Maia Martcheva. *An Introduction to Mathematical Epidemiology*. Springer New York, 2015. DOI: 10.1007/978-1-4899-7612-3.

[Met+53] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[Mil+04] Maja Miloslavsky et al. "Recurrent events analysis in the presence of time-dependent covariates and dependent censoring". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1 (2004), pp. 239–257. DOI: 10.1111/j.1467-9868.2004.00442.x.

[Mor+95] Christopher H. Morrell et al. "Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer". In: *Journal of the American Statistical Association* 90.429 (1995), pp. 45–53. DOI: 10.1080/01621459.1995.10476487.

[MSK] Memorian Sloan Kattering Cancer Center MSK. *PSA Doubling Time Calculator*. [Online]. URL: https://www.mskcc.org/nomograms/prostate/psa_doubling_time.

[Mul21] Claude P. Muller. "Do asymptomatic carriers of SARS-COV-2 transmit the virus?" In: *The Lancet Regional Health - Europe* 4 (2021), p. 100082. DOI: 10.1016/j.lanepe.2021.100082.

[Par+21] Nicola Parolini et al. "SUIHTER: a new mathematical model for COVID-19. Application to the analysis of the second epidemic outbreak in Italy". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 477.2253 (2021), p. 20210027. DOI: 10.1098/rspa.2021.0027.

[Pea+91] Jay D. Pearson et al. "Modeling longitudinal rates of change in prostate specific antigen during aging". In: *Proceedings of the Social Statistics Section*. American Statistical Association. 1991, pp. 580–585. URL: https://community.amstat.org/socialstatisticssection/home.

[Per+19] Ricardo Pereira Mestre et al. "Correlation between PSA kinetics and PSMA-PET in prostate cancer restaging: A meta-analysis". In: *European Journal of Clinical Investigation* 49.3 (2019), e13063. DOI: 10.1111/eci.13063.

[PH15] Sunhee Park and David J. Hendry. "Reassessing Schoenfeld residual tests of proportional hazards in political science event history analyses". In: *American Journal of Political Science* 59.4 (2015), pp. 1072–1087. DOI: 10.1111/ajps.12176.

[Pra+20] Mélanie Prague et al. "Population modeling of early COVID-19 epidemic dynamics in French regions and estimation of the lockdown impact on infection rate". In: *medRxiv* (2020). DOI: 10.1101/2020.04.21.20073536v2.

[PSW13] Nicholas G. Polson, James G. Scott, and Jesse Windle. "Bayesian inference for logistic models using Pólya-Gamma latent variables". In: *Journal of the American Statistical Association* 108.504 (2013), pp. 1339–1349. DOI: 10.1080/01621459.2013.829001.

[PT09] Cécile Proust-Lima and Jeremy M.G. Taylor. "Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach". In: *Biostatistics* 10.3 (2009), pp. 535–549. DOI: 10.1093/biostatistics/kxp009.

[PWP81] Ross L. Prentice, Benjamin J. Williams, and Arthur V. Peterson. "On the regression analysis of multivariate failure time data". In: *Biometrika* 68.2 (1981), pp. 373–379. DOI: 10.1093/biomet/68.2.373.

[R C] R Core Team. *R: A Language and Environment for Statistical Computing*. R package. [Online]. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html.

[RC10] Christian P. Robert and George Casella. "Metropolis-Hastings Algorithms". In: *Introducing Monte Carlo Methods with R*. Springer New York, 2010, pp. 167–197. DOI: 10.1007/978-1-4419-1576-4_6.

[RC19] Kimberlyn Roosa and Gerardo Chowell. "Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models". In: *Theoretical Biology and Medical Modelling* 16 (2019), pp. 1–15. DOI: 10.1186/s12976-018-0097-6.

[Reg+20] Naresh Regula et al. "Comparison of ^{68}Ga-PSMA-11 PET/CT with ^{11}C-acetate PET/CT in re-staging of prostate cancer relapse". In: *Scientific Reports* 10.1 (2020), pp. 1–10. DOI: 10.1038/s41598-020-61910-6.

[RF00] James M. Robins and Dianne M. Finkelstein. "Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests". In: *Biometrics* 56.3 (2000), pp. 779–788. DOI: 10.1111/j.0006-341X.2000.00779.x.

[RHB00] James M. Robins, Miguel Angel Hernan, and Babette Brumback. "Marginal Structural Models and Causal Inference in Epidemiology". In: *Epidemiology* 11.5 (2000), pp. 550–560. DOI: 10.1097/00001648-200009000-00011.

[Rob+] Xavier Robin et al. *pROC: an open-source package for R and S+ to analyze and compare ROC curves*. R package. [Online]. URL: https://https://cran.r-project.org/web/packages/pROC/index.html.

[Rob93] James M. Robins. "Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers". In: *Proceedings of the Biopharmaceutical Section*. American Statistical Association. 1993, pp. 24–33. URL: https://community.amstat.org/biop/home.

[RR92] James M. Robins and Andrea Rotnitzky. "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers". In: *AIDS Epidemiology: Methodological Issues*. Birkhäuser Boston, 1992, pp. 297–331. DOI: 10.1007/978-1-4757-1229-2_14.

[RR95] James M. Robins and Andrea Rotnitzky. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data". In: *Journal of the American Statistical Association* 90.429 (1995), pp. 122–129. DOI: 10.1080/01621459.1995.10476494.

[Sam13] Malcolm Sambridge. "A parallel tempering algorithm for probabilistic sampling and multimodal optimization". In: *Geophysical Journal International* 196.1 (2013), pp. 357–374. DOI: 10.1093/gji/ggt342.

[SC97] Elizabeth H. Slate and Kathleen A. Cronin. "Changepoint modeling of longitudinal PSA as a biomarker for prostate cancer". In: *Case Studies in Bayesian Statistics*. Springer New York, 1997, pp. 435–456. DOI: 10.1007/978-1-4612-2290-3_14.

[Sch82] David Schoenfeld. "Partial residuals for the proportional hazards regression model". In: *Biometrika* 69.1 (1982), pp. 239–241. DOI: 10.1093/biomet/69.1.239.

[She93] Ross M. Sheldon. "Chapter 4 - Markov Chains". In: *Introduction to Probability Models*. Fifth Edition. Academic Press, 1993, pp. 137–198. DOI: 10.1016/B978-0-12-598455-3.50007-6.

[ST10] Eric A. Suess and Bruce E. Trumbo. *Introduction to Probability Simulation and Gibbs Sampling with R*. Springer New York, 2010. DOI: 10.1007/978-0-387-68765-0.

[Sta+87] Thomas A. Stamey et al. "Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate". In: *New England Journal of Medicine* 317.15 (1987), pp. 909–916. DOI: 10.1056/NEJM198710083171501.

[TD04] Anastasios A. Tsiatis and Marie Davidian. "Joint modeling of longitudinal and time-to-event data: an overview". In: *Statistica Sinica* (2004), pp. 809–834. URL: https://www3.stat.sinica.edu.tw/statistica/oldpdf/A14n39.pdf.

[The+] Terry M. Therneau et al. *survival: Survival Analysis*. R package. URL: https://cran.r-project.org/web/packages/survival/index.html.

[TT17] Kathleen M.C. Tjørve and Even Tjørve. "The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family". In: *PLOS ONE* 12.6 (2017), pp. 1–17. DOI: 10.1371/journal.pone.0178691.

[Ver+16] Frederik A. Verburg et al. "Extent of disease in recurrent prostate cancer determined by [^{68}Ga]PSMA-HBED-CC PET/CT in relation to PSA levels, PSA doubling time and Gleason score". In: *European Journal of Nuclear Medicine and Molecular Imaging* 43 (2016), pp. 397–403. DOI: 10.1007/s00259-015-3240-1.

[Vic+09] Andrew J. Vickers et al. "Systematic review of pretreatment prostate-specific antigen velocity and doubling time as predictors for prostate cancer". In: *Journal of Clinical Oncology* 27.3 (2009), pp. 398–403. DOI: 10.1200/JCO.2008.18.1685.

[VM97] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer New York, 1997. DOI: 10.1007/978-1-4419-0300-6.

[WHO] World Health Organization WHO. *Coronavirus (COVID-19) Dashboard*. [Online]. URL: https://covid19.who.int.

[Wil+18] Sanne J.W. Willems et al. "Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator". In: *Statistical Methods in Medical Research* 27.2 (2018), pp. 323–335. DOI: 10.1177/0962280216628900.

[Xu87] Xiaoli Xu. "The biological foundation of the Gompertz model". In: *International Journal of Bio-Medical Computing* 20.1 (1987), pp. 35–39. DOI: 10.1016/0020-7101(87)90012-2.

[YS22] Sean Yiu and Li Su. "Joint calibrated estimation of inverse probability of treatment and censoring weights for marginal structural models". In: *Biometrics* 78.1 (2022), pp. 115–127. DOI: https://doi.org/10.1111/biom.13411. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13411. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13411.

Acknowledgements

When I was nineteen and had to choose what to study at university, I had a peculiar conversation with my dad. He was convinced that I could excel as a researcher and that mathematics was the right field for me. At that moment, the only thing I was absolutely sure about was that I didn't want to pursue a Ph.D. or a career in academic research. Today, ten years later, I wouldn't describe myself as a good researcher, but at least he was right in saying that I could manage to finish a Ph.D..

Completing this thesis has been a challenging yet rewarding journey, and it would not have been possible without the support and encouragement of many individuals and institutions.

Firstly, I would like to express my gratitude to my supervisor, Prof. Mauro Gasparini. Your guidance and support have been crucial in shaping this research and inspiring me throughout this process.

I am also grateful to GSK and my GSK supervisor, Marco Costantini, for their financial support, without which this research would not have been possible.

I am profoundly grateful to the members of my thesis committee, Prof. Michela Baccini and Prof. Pavel Mozgunov, for their time, effort, and thoughtful suggestions that have significantly improved the quality of this work.

My deepest thanks go to my co-supervisors, Enrico Bibbona and Gianluca Mastrantonio, whose guidance, expertise, and support have been invaluable throughout this journey. I truly appreciate the care you have shown for me, not only as a student but also on a personal level. I am grateful for the professional and personal moments we shared. Enrico, among many other things, thank you for taking the time to call me when I was in Belgium just to check that everything was fine (and for all the coffees you offered me). Gianluca, among many other things, thank you for always being online to patiently listen to my daily complaints (and for all the coffees you offered me).

A special thanks to Prof. Stijn Vansteelandt and all the members of his statistical group at Ghent University for their hospitality. The period I spent in Ghent was extremely instructive and full of stimulating discussions and nice talks. It was a fantastic experience, and each of you made me feel at home.

Thanks to Andrea Callegaro, who made this Belgian experience possible and always found time for a call when I needed it.

I would also like to thank my colleagues and friends at Politecnico di Torino and Ghent University. The stimulating discussions, collaborative projects, and, in particular, moral support have been essential in keeping me motivated and focused. I don't need to mention each of you by name, as you know how helpful your contributions were. Long calls and messages to stay updated even from a distance, late-evening sports to recover from difficult days (with the special attitude that only dream teams have), shared meals and food at each moment of the day, long discussions over big cups of tea, karaoke and board-games nights, walks with ice cream and many complaints, and much more – your support has been invaluable.

Thanks to all the new statistician friends I met at conferences, who encountered the most stressed and anxious version of me but still chose to be part of this journey.

Thanks to the friends who were already around before and are still part of my life. To all members of the Giuggiole team, thank you for your patience and support and for trying to understand, time after time, what I was doing as a Ph.D. student.

A small but sincere thank to who took care of the bibliography and the layout of this thesis (but most importantly, of my mental health during the final stages).

On a personal note, I am deeply thankful to my parents, Silvia and Giuseppe. Even if it was not always clear to you what I was doing, what I was studying, what I was stressed for, and what I was complaining about, your (almost) endless patience, encouragement, and belief in me were essential!

Un grazie, in Italiano, anche ai Nonni che non hanno mancato di comunicare, con tanto orgoglio, tutti i miei progressi a tutto il personale dell'Esselunga e del San Luigi.

Finally, my deepest thanks go to my sister Carola, with whom I shared my home office during the long pandemic. I know you see me as the big sister who always knows how to handle things and can help you with your problems, but you are the same for me – just in a kinder and more smiling version.