Statistical modelling of time-varying covariates for survival data

Spreafico, M.

Introduction

Survival analysis [97, 106, 202] is an important field of statistics dealing with time-to-event (or life-time) data. In life-time data the outcome variable of interest, named *survival time*, is the amount of time elapsed from a so-called origin/initiating event until an event of interest. Many examples exist in several research fields: time from diagnosis of disease until death in medicine, time to failure of a machine part in engineering, or duration of unemployment in socio-economic sciences. A key characteristic of survival data is that they are generally *partially observed*, coming as a mixture of complete and incomplete observations. Some individuals might not have experienced the event of interest at the end of the study period or have dropped out before the event has occurred. It is only known that the event had not occurred before the last observation time but the exact event time is unknown. This type of mechanism is called *right-censoring*. Whether studying a specific event or a sequence of events, special statistical methodologies are required to handle this particular type of partially observed data, as *censoring* complicates all technical issues involved in life-time analyses.

The work carried out in this thesis is motivated by specific medical questions. For example in chronic diseases or cancers, survival models can be used to investigate if a patient's age, gender, medical treatment, or other covariates are associated to the risk of death. Typically, the term covariate refers to *time-fixed* predictive or explanatory variable, whose value does not change over time (e.g., demographic or baseline information). However, medical follow-ups are characterized by *time-varying covariates* (i.e., attributes that may have different values at different time points), such as drugs intake, treatment doses, biomarkers or toxicity, or by *repeated events* occurring during the study, like office visits, subsequent drug consumption or hospital admissions. These processes that change or re-occur over time are of great interest because the way their dynamic patterns evolve may affect patient's health status and disease progression.

Due to the complexity of these phenomena, a *piecewise-constant* or *fixed-baseline* approach is usually preferred in the clinical literature, discarding their dynamic and/or temporal components. In this way, the information that these processes can provide if the association between time-varying and time-event data is properly captured is completely lost. Complex mathematical methods are therefore required to model disease evolution and characterise its relationship to the dynamic nature of time-dependent features.

The current thesis arises in cross-sectional fields of biostatistics and healthcare research and focuses on developing mathematical and statistical methods to represent time-varying processes from complex raw data, and model them within the context of time-to-event analysis. The main purpose is to enrich the knowledge available for modelling survival

with relevant features related to the dynamic characteristics of interest. These aspects are rarely addressed in the literature and may provide new insights for medical research, representing a challenge of both clinical relevance and statistical interest. This work focuses on time-varying processes and, more specifically, examines (i) dynamic *representation* and (ii) *modelling* in a time-to-event setting.

In terms of *representation*, the main issue consists on identifying appropriate dynamic characterizations of the processes under study. Here, several levels of complexities must be considered. On one hand, producing models that are suitable for dealing with complex data is not straightforward. A huge amount of data-integration and preprocessing work is needed to make data suitable for the statistical analysis preserving clinical interpretability. On the other hand, when defining the mathematical formulation, the nature of the processes and aspects such as temporal dynamics, categorical levels or recurring events must be taken into account. Different methodologies are therefore proposed throughout this thesis, including:

- complex data integration to define novel pooled or longitudinal representations related to time-dependent covariates (Chapters 1, 2, 5);
- recurrent events modelling and point processes theory to retrieve the trajectories of compensators related to appropriate stochastic processes for recurrent events (Chapter 3);
- functional data analysis techniques to reconstruct features able to incorporate trends and variations in the evolution of the processes as continuous smoothed functions of time (Chapters 3, 4);
- latent Markov models and compositional data analysis to model latent disease evolution on the basis of interval-based categorical observations (Chapter 6);
- direct acyclic graphs to identify all possible confounders and their relation with the time-dependent exposure under study and then engage a causal inference paradigm (Chapter 7).

At this stage, the main challenge is to best represent the time-varying characteristics of interest by managing the complex trade-off between clinical interpretability and mathematical formulation.

In terms of time-to-event *modelling*, innovative statistical models for identifying and quantifying the association between time-varying processes and patient survival are proposed. In medical statistics the Cox proportional hazards model [46] has been widely used for survival data due to its flexibility. It has also been extended to account for time-dependent covariates [202, 97] using piecewise-constant values among different time measurements. Discarding the continuous nature of the process underlying the data, this approach leads to biased results and fails to account for possible measurement errors [16]. Therefore, the main purpose during the survival *modelling* phase is to develop advanced versions of Cox-type methods capable of handling dynamic covariates, while preserving clinical

interpretability. Depending on the context of the study, different approaches are then proposed, such as:

- a joint modelling to simultaneously analyse longitudinal and time-to-event data through appropriate mixed-effect and Cox-type models (Chapter 2);
- dimensionality reduction techniques for functional data in order to incorporate dynamic predictors into advanced version of traditional Cox regression models (Chapters 3, 4);
- Cox-type marginal structural models to assess the causal effects of a joint-exposure on survival outcomes, in presence of time-varying confounders (Chapters 7).

The main purpose is to add to the current literature relevant survival models which are able to incorporate dynamic information usually discarded by standard approaches.

This work has an impact on the community of researchers in mathematics and statistics, but it provides also useful tools to support doctors and clinicians in their daily work. All research topics are motivated by specific clinical questions related to two different medical domains, corresponding to the two main parts of the thesis: cardiological (Part I) and oncological (Part II) patients.

The identification of dynamic representations able to reflect variability and differences among patients may improve patients' profiling and tailor their therapies. This can lead to new knowledge for both general guidelines and personalised treatments, and make the pathway of patients through the healthcare system more efficient and effective. Therefore, the development of novel methodologies capable of extracting additional information to enrich survival models may represent a significant step forward in the definition of new customized and flexible monitoring tools, which could then be applied to the study of different pathologies characterised by complex data sources.

The remainder of this Chapter is composed as follows. Section I.1 introduces basic concepts of survival analysis. Section I.2 presents the motivating epidemiological and clinical frameworks. Section I.3 gives an overview of the remaining Chapters of this thesis.

I.1. Basics of Survival Analysis

This section aims at providing notation for the rest of the thesis. A first step in understanding survival analysis [97, 106, 202] is in understanding the *partially-observed* time-to-event data it has to deal with.
Let T^* be the non-negative random variable denoting the true event time, i.e., the amount of time elapsed from the origin event until an event of interest. If a patient dropped out of the study early or the study ended before the event of interest occurred, or another event occurred, the event time may not be observed and right-censoring occurs. Let C be a random variable that denotes the time for the censoring mechanism (i.e., the last time a subject was observed in the study). The time-to-event information observed for

an individual is given by the pair (T, D), where $T = \min(T^*, C)$ is the *survival time* and $D = \mathbb{1}(T \leq C)$ is the *event indicator*, $D = 1$ if the true event time was observed or $D = 0$ if censored. The event time T^* and the right-censoring time C are usually assumed to be independent.

To study the distribution of the survival time T different quantities are of interest. The *survival function* (or survival curve) $S(t) = \Pr(T > t)$ at time t is equal to the probability of being event-free at time t. It is a non-increasing function with $S(0) = 1$ because everybody is event-free at the time origin, and as t gets large as $S(t)$ tends to 0 because everybody eventually experiences the event of interest. The *hazard function*, i.e., the instantaneous risk of failure at time t, conditional on survival to that time, is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

The hazard function can be used to express:

- the *survival function*:

$$S(t) = \exp\left\{ - \int_0^t h(u)du \right\},$$

  which can be estimated by using a non-parametric method called the product limit estimator better known as the *Kaplan-Meier* estimator [99]

$$\hat{S}(t) = \prod_{j:\, t_j^* \leq t} \left(1 - \frac{d_j}{n_j}\right);$$

- the *cumulative hazard function* representing the total accumulated risk of experiencing the event of interest that has been gained by progressing to time t:

$$H(t) = \int_0^t h(u)du,$$

  which can be estimated by using the *Nelson-Aalen* estimator [145, 146, 1]

$$\hat{H}(t) = \sum_{j:\, t_j^* \leq t} \frac{d_j}{n_j}.$$

In both estimators $\hat{S}(t)$ and $\hat{H}(t)$, $0 < t_1^* < t_2^* < \cdots < t_J^* < \infty$ denote the observed ordered true event times with J equal to the total number of events, d_j and n_j denote the number of events and the number of individual still at risk at time t_j^*, respectively. Figure I.1 shows an example of time-to-event data (left panel) and the corresponding estimated survival and cumulative hazard curves (central and right panels, respectively). Subjects 2, 5, 8 and 12 are right-censored. Both curve have steps at event times (red points) and remain unchanged at censoring times (light-blue diamonds). The censoring times however affect the size of the jumps the curves make.

The main goal of survival studies is to estimate the hazard function and to assess how the covariates affect it. The most widely used model to study the effect of a covariate

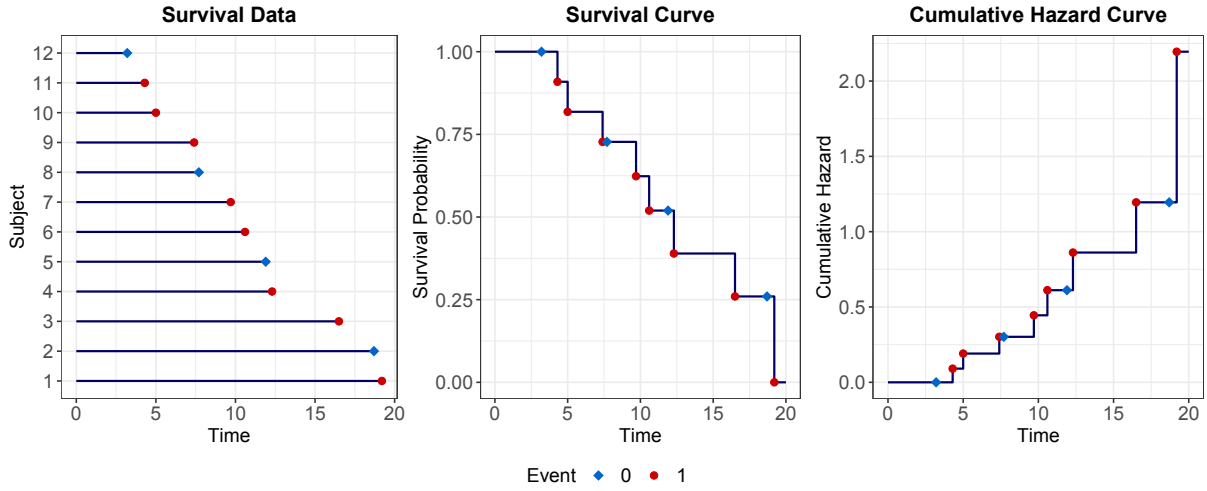

Figure I.1. Left panel: time-to-event data for 12 subjects (*light-blue diamonds*: censored subjects; *red points*: event subjects). Central panel: Kaplan-Meier survival estimate. Right panel: Nelson-Aalen cumulative hazard estimate.

vector $\boldsymbol{\omega}$ on the survival is the Cox proportional-hazard regression [46]. It is based on the proportional hazards assumption stating the effects of the covariates are multiplicatively related to the hazard, defined as:

$$h\left(t|\boldsymbol{\omega}\right) = h_0(t)\exp\left\{\boldsymbol{\theta}^T\boldsymbol{\omega}\right\},$$

where $h_0(t)$ is the unspecified non-negative baseline hazard function and $\boldsymbol{\theta}$ is the vectors of regression coefficients. Inference for coefficients $\boldsymbol{\theta}$ is based on maximizing the *partial likelihood* [46]:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} \frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{\omega}_{(j)}\right)}{\sum_{i\in R(t_j^*)}\exp\left(\boldsymbol{\theta}^T\boldsymbol{\omega}_i\right)}$$

where $0 < t_1^* < t_2^* < \cdots < t_J^* < \infty$ are the observed ordered event times, $\boldsymbol{\omega}_{(j)}$ denotes the covariates of the individual who experiences the event at time t_j^*, $\boldsymbol{\omega}_i$ is the covariate vector of individual i and $R(t_j^*)$ denotes the set of individuals still at risk at time t_j^*. The baseline hazard $h_0(t)$ can then be estimated by *Breslow* estimator [32].

For each covariate l, the quantity $\exp(\hat{\theta}_l) = HR_l$ is called Hazard Ratio (HR). For a categorical explanatory variable, the HR represents the ratio between the predicted hazard for a member of one group and for a member of the reference group, by holding everything else constant. For a continuous explanatory variable, the same interpretation applies to a 1-unit difference. In particular, a $HR_l < 1$ indicates a reduction in the hazard function (i.e., an increase in the survival), meaning that the l-th covariate is a *protective* factor for the time to the event of interest; whereas the opposite ($HR_l > 1$) indicates that the l-th covariate is a *risk* factor.

The Cox model has also been extended to account for a covariate vector $\boldsymbol{\omega}(t)$ which can change values during follow-up [202, 97]. Since time-dependent observations are only available at the times of measurements, the Time-Varying covariate Cox Model (TVCM) uses the last-observation-carried-forward (LOCF) approach [206]: between two subsequent

observations, the value of the time-varying covariate is kept constant at the last observed value. Under TVCM, the hazard function is

$$h\left(t|\boldsymbol{\omega}(t)\right) = h_0(t)\exp\left\{\boldsymbol{\theta}^T\boldsymbol{\omega}(t)\right\}.$$

The partial likelihood is defined similarly to the model with only time-fixed covariates [202, 97], considering piecewise-constant values for the time-dependent covariates.

I.2. Epidemiological and clinical framework

As mentioned before, the research topics of this thesis are related to two different applications: (i) the study of pharmacotherapy in patients with heart failure and (ii) the investigation of chemotherapy treatment in patients with osteosarcoma. The motivating clinical research issues for both cases are now introduced.

I.2.1. Pharmacoepidemiology in Heart Failure

Pharmacoepidemiology is the study of the use and the effects of drugs in large numbers of people though the application of epidemiological methods [194, 195]. A modern definition of pharmacoepidemiology [220] is

"the study of the use and effects/side-effects of drugs in large numbers of people with the purpose of supporting the rational and cost-effective use of drugs in the population thereby improving health outcomes".

The investigation about quantification, understanding and evaluation of the processes of prescribing, dispensing and consumption of medicines and their effect on patients' clinical courses refers to a branch of pharmacoepidemiology known as Drug Utilization Research (DUR) [53]. As defined by the World Health Organization in 1977 [220], DUR consists in the

"marketing, distribution, prescription, and use of a drug in the society, with special emphasis on the resulting medical, social and economic consequences".

The ultimate goal of DUR is hence to identify and communicate the proper use of drugs to patients, combining researches which belong to the medical, economical and social fields.

In DUR and pharmacotherapy, the achievement of a certain level of medication intake or adherence is an important component of patient's care. According to the taxonomy introduced in [212], *adherence* to medication is defined as the process by which patients take their medication as prescribed. In the past decade there has been substantial growth in clinical research focused on adherence to medication, partly owing to the increasing awareness of the problem and partly due to the pervasiveness of non-adherence behaviours among patients [212]. Individual patient's adherence is usually reported as a percentage of the actual medication taken over a defined period of time [27]. Poor adherence to

medication regimens accounts for substantial worsening of disease, death and increased health care costs [147]. In particular, in long-term therapies poor adherence severely compromises the effectiveness of treatment, representing a critical issue both for quality of life and health economics [219].

Long-term therapies are typical in chronic diseases, such as Heart Failure (HF). HF is a major and growing public health issue, characterized by high costs, steep morbidity and mortality rates [129]. HF is widespread all over the world, especially among people over 65 years, with a prevalence of 1–2% in Western countries and an incidence from 5 to 10 per 1000 persons per year [150]. In particular, in Italy about 80,000 new cases per year are recorded [131] and it is the second cause of hospitalization, after vaginal delivery. This complex clinical syndrome is characterized by structural or functional cardiac disorders that impair the ability of one or both ventricles to fill with or eject blood [87]. It may be provoked by several different cause, such as myocardial ischaemia, high blood pressure, cardiomyopathies, valvular heart disease, pulmonary hypertension or congenital heart disease [150]. Due to HF, organs and tissues receive insufficient quantities of oxygen and nutrients for their metabolic needs, and there is an accumulation of excess fluid in the lungs and tissues [143]. This condition can worsen to the point of acute pulmonary oedema and death. According to data from different studies conducted in America and Europe, 30-day, 1-year, and 5-year mortality are around 10% to 20%, 30%, and 65% respectively [150].

Patients hospitalized for HF are at high risk for all-cause re-hospitalization, with a 1-month readmission rate of 25%. Self-care in HF comprises treatment adherence and health maintenance behaviours [75]. HF patients should learn to take medications as prescribed, stay physically active, restrict sodium intake, get vaccinations and understand how to monitor for signs of worsening HF [75]. Therapeutic and pharmacological interventions in HF patients aim at reducing symptoms, morbidity and mortality. Depending on the different symptoms, the following pharmacological treatments have been established as disease-modifying drugs of routine use in HF treatment: Angiotensin-Converting Enzyme (ACE) inhibitors, Angiotensin II Receptor Blockers (ARB – as an alternative for people who cannot tolerate ACE), Beta Blockers (BB), Anti-Aldosterone agents (AA) and diuretics [139, 222, 221, 154, 75]. Different studies showed that a proper and monitored drug intake in HF patients could improve their clinical status, functional capacity and quality of life, prevent hospital admission and reduce mortality [154]. However, it is well known that adherence in HF is low and not satisfactory [179], even few months after the first hospital discharge for HF. Poor adherence to medications leads to increased HF exacerbations, reduced physical function, higher risk for hospital readmission and death [154], representing a significant problem in HF management in both healthy and economic terms.

In the last decades, secondary or administrative databases have increasingly been used in the pharmacotherapy field [14], becoming one of the most employed sources to evaluate adherence to medication [102]. Although administrative data are mainly collected for managerial and economic purposes [89], their use for clinical and epidemiological purposes has become an accepted practice [90] as they are particularly suitable for investigating different areas, such as profile of drug uses [45]. This requires intensive computational

effort to link different data sources (e.g., drug purchases, death registry, hospitalisation records) in order to create usable databases. According to the state of the art, adherence to medication is usually modelled by a numerical variable representing the percentage of the actual drug taken over a pre-defined period of time [27]. This approach does not consider *changes in drugs consumption over time* or the occurrence of *re-hospitalizations*. Moreover, the most used adherence measures [14, 102] refer to monotherapy medication, although chronic HF patients usually undergo a *polypharmacy treatment* (i.e., the simultaneous use of different medications). Given the clinical relevance of these aspects and the impact they may have on patients' survival, the development of new methodologies to overcome these problems is a challenge for both clinical research and statistical modelling. This topic is discussed in Part I of this thesis.

I.2.2. Chemotherapy in Osteosarcoma

Osteosarcoma is a malignant bone tumour mainly affecting children and young adults. Although osteosarcoma is the most common primary malignant bone cancer, it is a rare disease and has an annual incidence of 3-4 patients per million [185]. Osteosarcoma can occur in any bone but it is often localized in the extremities: the most common primary sites are the distal femur, the proximal tibia, and the proximal humerus, with $> 50\%$ originating around the knee [166, 15]. Local pain, followed by localized swelling and limitation of joint movement, are the typical signs and symptoms of osteosarcoma [166].

Osteosarcoma treatment typically involves surgery and chemotherapy. The goal of surgery is the complete tumour removal. Different surgical techniques are available and the surgeon must choose the most appropriate for each individual, taking into account several factors such as the size of the tumour and its location, as well as the influence the surgery will have on the patient's daily life. In case of unresectable tumours or of microscopic residual tumour foci following attempted resection, recent research also suggests radiotherapy in addition to standard therapy [166]. In modern treatment schedules, chemotherapy to kill cancer cells is usually a combination of doxoubicine (DOX) and cisplatin (CDDP), with or without high-dose methotrexate and/or ifosfamide and/or etoposide. DOX and CDDP are considered the most active agents against osteosarcoma, but the ideal combination remains to be defined [166, 15]. Since the extent of histological response to pre-operative chemotherapy (i.e., the improvement in the appearance of microscopic tissue specimens) represents the strongest prognostic factor of survival known so far in osteosarcoma [31], most current protocols include a period of pre-operative (neoadjuvant) chemotherapy. Post-operative (adjuvant) chemotherapy is then used to kill any cancer cells that might remain after surgery.

Multidisciplinary management including neoadjuvant and adjuvant chemotherapy with aggressive surgical resection [166] or intensified chemotherapy has improved clinical outcomes although the overall 5-year survival rate has remained unchanged in the last 40

years at 60-70% [15]. The impact of chemotherapy dose modification on patients' survival is still unclear [111]. In cancer trials the relationship between chemotherapy dose and clinical efficacy outcomes is difficult to analyse due to the presence of negative feedback between exposure to cytotoxic drugs and other aspects, such as latent accumulation of chemotherapy-induced toxicity. Patients may develop different types of toxic side effects, ranging in severity from minor, asymptomatic changes to life-threatening injuries or death [204]. Depending on patients' treatment history or development of toxicity, biomarkers values change and chemotherapy treatment is modified by delaying a course or reducing the dose intensity. Being at the same time risk factors for mortality and predictors of future exposure levels, toxicities are time-dependent confounders for the effect of chemotherapy on patient's survival [112].

Due to the complexity of longitudinal chemotherapy data, the ways chemotherapy doses and toxicities are accounted for into predictive models in literature and current practice for cancer research is far from being informative as it may be. Chemotherapy is usually modelled by different allocated regimens, i.e., by Intention-To-Treat (ITT) analysis [70]. This means that protocol deviations or changes in drug intake over time are not considered in the analysis [110]. Toxicities are usually incorporated as summary indexes (e.g., maximum toxicity over time, maximum grade among events, or weighted sums of individual toxic effects) discarding substantial amount of information (e.g., isolated vs repeated events, single vs multiple episodes, longer-lasting lower-grade toxicities, toxic events timing). As neglecting the time component may give an inaccurate depiction of toxicity and chemotherapy regimen intensity, characterisation of both aspects is of interest to patients and clinicians engaged in shared decision making about a treatment strategy. The development of models and methods able to deal with all these peculiar aspects is hence of statistical interest and of clinical relevance. This topic is discussed in Part II of the this thesis.

I.3. Overview of the thesis

The current thesis aims at developing mathematical and statistical methods to properly *represent* time-dependent processes and *modelling* them within the context of time-to-event analysis by means of appropriate Cox-type survival models.

Part I *"Pharmacoepidemiology in Heart Failure"* focuses on methods for representing drug consumption, adherence to medications or re-hospitalization events exploiting administrative databases, and modelling their effect on long-term survival in HF patients. Administrative data from *Friuli Venezia Giulia* and *Lombardia* [164] regions in Italy are analysed. In particular, records from Hospital Discharge Charts (i.e., admission to hospital), Public Drug Distribution Systems (i.e., drugs purchases) and Registries of Deaths are considered.

In Chapter 1 we investigate patients' adherence to disease-modifying therapies and the prognostic impact on survival, exploiting administrative databases of the *Friuli Venezia*

Giulia Italian region. A novel method to represent adherence to polypharmacy, i.e., the Patient Adherence Index (PAI), is proposed as the ratio between the number of drugs to which a patient is adherent and the number of purchased drugs. Taking advantage of the developed index, the effect of adherence to polypharmacy on patients survival is then investigated through Cox regression model, adjusting for patient-specific characteristics. Although PAI is still a time-fixed covariate, this study requires complex data integration procedures among different data sources, representing a first step forward in the pharmacoepidemiology context for HF patients as few data on polypharmacy adherence exist in a real-world setting. The content of this chapter is based upon:

- **M. Spreafico**, F. Gasperoni, G. Barbati, F. Ieva, A. Scagnetto, L. Zanier, A. Iorio, G. Sinagra and A. Di Lenarda. *Adherence to Disease-Modifying Therapy in Patients Hospitalized for HF: Findings from a Community-Based Study.* American Journal of Cardiovascular Drugs, 20:179–190, 2020 [187].

In Chapter 2 we propose an innovative method to represent adherence to medication as time-varying covariate and to investigate its dynamic effect on patients' survival using a joint modelling framework. Two different longitudinal representations are introduced: a continuous time-dependent variable, which indicated the cumulative months covered by drug consumption up to time t, and a dichotomous time-dependent variable, which indicates if the patient is adherent to the therapy at time t. The development of (generalized) mixed effect models for these longitudinal processes joint with Cox-type regression model for time-to-death allows to capture the interaction among processes over time, representing a more realistic and informative approach with respect to the commonly used time-fixed measures. Administrative databases of the *Lombardia* Italian region provided by *Regione Lombardia - Healthcare Division* [164] are analysed. The content of this chapter is based upon the following publication:

- **M. Spreafico** and F. Ieva. *Dynamic monitoring of the effects of adherence to medication on survival in heart failure patients: A joint modeling approach exploiting time-varying covariates.* Biometrical Journal, 63(2):305–322, 2021 [188].

Chapter 3 concerns the development of a new methodology to extract and summarize information from trajectories of compensators of suitable marked point processes for recurrent events intended as functional data. The developed methodology involves database integration, Functional Data Analysis (FDA) and marked point process modelling of critical events of interest, i.e., drugs purchases and re-hospitalizations. First, the functional trajectories (i.e., the compensators of such processes, which may represent the rate at which events happen) are retrieved by means of FDA theory. This new information is then included into a predictive Cox-type model, exploiting Functional Principal Component Analysis (FPCA) techniques. The introduction of this novel way to account for dynamic processes allows for modelling self-exciting behaviours, for which the occurrence of events in the past increases the probability of a new event, including a large piece of information about patient's clinical history contained in the administrative data. The developed approach is able to take into account the fact that HF patients usually experience several re-hospitalizations and consume different types of drugs at the same time, representing a novelty for clinical and pharmacological research in the direction of properly

treating multimorbidity and polypharmacy. Administrative databases of the *Lombardia* Italian region [164] are analysed. This chapter is based on the following publication:

- **M. Spreafico** and F. Ieva. *Functional modeling of recurrent events on time-to-event processes*. Biometrical Journal, 63(5):948–967, 2021 [189].

Part II *"Chemotherapy in Osteosarcoma"* focuses on methods to represent and model chemotherapy treatment and related effects in cancer patients, such as dose modifications, biomarkers changes, toxicities evolution over time and their associations with survival. Clinical data from randomized trials funded by the Medical Research Council (MRC) (`https://www.ukri.org/councils/mrc/`) and the European Organisation for Research and Treatment of Cancer (EORTC) (`https://www.eortc.org`) for patients with high-grade osteosarcoma are analysed.

In Chapter 4 we propose a Functional covariate Cox Model (FunCM) to study the association between time-varying processes and time-to-death outcome. FunCM first exploits FDA techniques to represent time-varying processes in terms of functional data. Then, information related to the evolution of the functions over time is incorporated into functional regression models for survival data through FPCA. FunCM is compared to a standard TVCM, commonly used despite its limiting assumptions that covariate values are piecewise-constant in time and measured without errors. Data from MRC BO06/EORTC 80931 randomised controlled trial [119] are analysed. Time-varying covariates related to alkaline phosphatase levels, white blood cell counts and chemotherapy dose during treatment are investigated. The proposed method allows to detect differences between patients with different biomarkers and treatment evolutions, and to include this information in the survival model. The content of this chapter is based on the following work:

- **M. Spreafico**, F. Ieva and M. Fiocco. *Modelling time-varying covariates effect on survival via Functional Data Analysis: application to the MRC BO06 trial in osteosarcoma*. Statistical Methods & Applications, 2022 [192]. `https://doi.org/10.1007/s10260-022-00647-0`

Chapters 5 and 6 focus on the methodological aspects concerning a proper representation of the overall toxicity burden over time, still lacking in the medical literature due to the complex nature of both chemotherapy protocol and data. In both cases, data from the MRC BO06/EORTC 80931 randomized clinical trial [119] are analysed. In Chapter 5 we exploit complex database processing and aggregation methods to introduce two innovative longitudinal representations of Multiple Overall Toxicity (MOTox), a continuous mean-max score and a dichotomous one. These new representations are then used to investigate the evolution of high-MOTox over treatment through the implementation of cycle-specific multivariable logistic regression models adjusted for previous toxicity levels and patient's characteristics. Although this approach represents a flexible method for quantifying the individual evolution of overall toxicity in cancer patients compared to traditional indexes, it discards the categorical nature of the observed toxic grades. For this reason, in Chapter 6 we propose a new taxonomy based on latent Markov models with covariates and compositional data theory to (i) represent the overall toxicity as the latent process that

affects the distribution of the observed response variables (i.e., the interval-based categorical toxic levels), (ii) identify different states of Latent Overall Toxicity (LOTox) burden, and (iii) model the personalized *longitudinal LOTox profiles* representing the probability over time of being in a specific LOTox state or the relative risk with respect to a reference "good" toxic condition. Together, absolute probabilities and relative risks provide a full picture of the individual LOTox dynamics during treatment, which may be considered as a proxy for patient's quality of life and used to describe patient's response to therapy over cycles in terms of toxic side effects. Chapter 5 is based upon the following publication:

- **M. Spreafico**, F. Ieva, F. Arlati, F. Capello, F. Fatone, F. Fedeli, G. Genalti, J. Anninga, H. Gelderblom and M. Fiocco. *Novel longitudinal Multiple Overall Toxicity (MOTox) score to quantify adverse events experienced by patients during chemotherapy treatment: a retrospective analysis of the MRC BO06 trial in osteosarcoma*. BMJ Open, 11(12):e053456, 2021 [190].

Chapter 6 is extracted and extended from the following work:

- **M. Spreafico**, F. Ieva and M. Fiocco. *Longitudinal Latent Overall Toxicity (LOTox) profiles in osteosarcoma: a new taxonomy based on latent Markov models*. arXiv:2107.12863, 2021 [191]. *[Submitted]*

In Chapter 7 we introduce marginal structural models in combination with Inverse-Probability-of-Treatment Weighted estimators to model the causal effects of chemotherapy intensity exposure on Event-Free Survival (EFS) in presence of time-dependent confounders. Statistical and clinical expertises are merged to propose a suitable characterisation of the causal structure of the chemotherapy data through the introduction of appropriate direct acyclic graphs that identify all possible (time-dependent) confounders (i.e., toxicities and other individual characteristics) and their relationships with exposure and EFS outcome. Data from the control arms of European Osteosarcoma Intergroup studies MRC BO03/EORTC 80861 [120] and MRC BO06/EORTC 80931 [119] for patients with osteosarcoma are analysed. Since drug administration is longitudinal while only the most severe side-effects are recorded, the analysis of such mixed longitudinal/non-longitudinal data requires both an original analytical strategy and an unconventional model formulation. The main contribution of this chapter is the presentation of an all-round analysis of complex chemotherapy data, with tutorial-like explanations of the difficulties encountered and the problem-solving strategies deployed. The content of this chapter is based on the following work:

- **M. Spreafico**, C. Spitoni, C. Lancia, F. Ieva and M. Fiocco. *Causal effects of chemotherapy regimen intensity on survival outcome in osteosarcoma patients through Marginal Structural Cox Models*, 2022. *[Submitted]*

In the final Conclusions we summarise the contributions and achievements of this work from both a statistical and clinical point of view, identifying the added values of the entire thesis from a global perspective.

Codes to perform the analysis is written in R software environment [161] and are available on my personal Github repository (`https://github.com/mspreafico`) or as supplementary material of the relative published papers.